

# CAUSAL DISCOVERY IN SOCIAL WEATHER SYSTEM

by

**Jiexiao He**

Bachelor, Wuhan University, 2015

Submitted to the Graduate Faculty of  
the Department of Informatics and Networked Systems in partial  
fulfillment

of the requirements for the degree of

**Master of Science**

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH  
SCHOOL OF COMPUTING AND INFORMATION

This thesis was presented

by

Jiexiao He

It was defended on

March 18th 2019

and approved by

Vladimir I. Zadorozhny, Ph.D., Professor, University of Pittsburgh

Marek J. Druzdzel, Ph.D., Professor, University of Pittsburgh

Paul Munro, Ph.D., Associate Professor, University of Pittsburgh

Thesis Advisor: Vladimir I. Zadorozhny, Ph.D., Professor, University of Pittsburgh

Copyright © by Jiexiao He  
2019

# CAUSAL DISCOVERY IN SOCIAL WEATHER SYSTEM

Jiexiao He, M.S.

University of Pittsburgh, 2019

In this thesis, we explore relationships between social variables that can be used to assess social events and conditions (social weather). Social weather may have hidden causes, and social events may be related to several variables distributed over diverse data sources. We built an infrastructure integrating data sources using the World Bank, stock market, happiness, terrorism, and Internet usage in the US. We performed data cleaning, data normalization and data aggregation and implemented our data store using Influx DB; we used Grafana for visual exploration of the integrated database.

The integrated data store allowed us to explore data trends and relationships among the social variables. In particular, we explored causal links among large amount of social variables using the Tetrad framework. We applied several causal discovery algorithms to generate a support matrix and used the voting approach to find strong causal relationships. We demonstrated that our causal discovery results are consistent with a well-known economic model.

**Keywords:** Time-series System, Causal Discovery, Majority Vote.

## TABLE OF CONTENTS

<b>PREFACE</b> . . . . .	x
<b>1.0 INTRODUCTION</b> . . . . .	1
<b>2.0 LITERATURE REVIEW</b> . . . . .	5
<b>3.0 TIME-SERIES DATABASE</b> . . . . .	8
3.1 DATA COLLECTION AND DATA CLEANING . . . . .	9
3.1.1 Data collection . . . . .	9
3.1.2 Data description . . . . .	9
3.1.3 Data cleaning . . . . .	10
3.1.4 Front End Design . . . . .	12
3.1.4.1 Front-End Design for Stock market . . . . .	13
3.1.4.2 Front-End Design for the World Bank . . . . .	16
3.1.4.3 Terrorism visualization . . . . .	17
3.1.4.4 Front-End Design for Internet Usage . . . . .	18
3.1.4.5 Front-End Design for Happiness . . . . .	20
<b>4.0 CAUSAL DISCOVERY AND SUPPORT MATRIX</b> . . . . .	21
<b>5.0 MATHEMATICAL FOUNDATION</b> . . . . .	24
5.0.1 DAG . . . . .	24
5.0.2 d-separate . . . . .	24
5.0.3 Causal Markov Condition . . . . .	24
5.0.4 Markov Equivalence . . . . .	25
5.0.5 Skeleton . . . . .	25
5.0.6 CPDAG . . . . .	25

5.0.7 Bayesian network . . . . .	26
<b>6.0 CAUSAL DISCOVERY ALGORITHMS . . . . .</b>	<b>27</b>
6.0.1 Correlation . . . . .	27
6.0.2 Fast Greedy Equivalence Search . . . . .	28
6.0.3 Greedy Fast Causal Inference (GFCI) and Fast Causal Inference (FCI)	32
6.0.4 Fast Adjacency Search(FAS) and PC Algorithm . . . . .	34
6.0.4.1 Majority vote . . . . .	37
<b>6.1 EVALUATE . . . . .</b>	<b>45</b>
6.1.1 Reference data and Model Behavior . . . . .	45
6.1.2 Jaccard index . . . . .	48
<b>7.0 CONCLUSIONS . . . . .</b>	<b>51</b>
<b>BIBLIOGRAPHY . . . . .</b>	<b>53</b>

## LIST OF TABLES

1	Definition of Lines for Support Matrix . . . . .	22
2	Table for rules of data . . . . .	31
3	Pairs of nodes by algorithm support . . . . .	43
4	Pairs of nodes by algorithm support . . . . .	44
5	Table for reference model's variables table . . . . .	47
6	Table for reference majority vote results. . . . .	49
7	Table for reference majority vote results. . . . .	50

## LIST OF FIGURES

1	Example of social variables cause and reasons . . . . .	2
2	Example of unknowable social variables causes and results . . . . .	2
3	Design flow map . . . . .	7
4	Summary for data aggregation . . . . .	11
5	The database used a measurement aggregated all datasets . . . . .	11
6	The visualization of Social Weather System in Grafana . . . . .	12
7	Main Page for Social Weather . . . . .	13
8	Stock Page - Stocks . . . . .	14
9	3D graph for Google, Amazon and Microsoft . . . . .	15
10	Stock Marketing by comparison . . . . .	15
11	General Data for world Bank . . . . .	16
12	Comparison of Internet Usage vs Mobile Subscription . . . . .	16
13	Comparison of Internet Usage vs Electric Consumption . . . . .	17
14	Front End Design for Terrorism Web Page . . . . .	18
15	Comparison of Internet with SP500 . . . . .	19
16	Comparison of Internet with DJIA . . . . .	19
17	Comparison of happiness and Internet . . . . .	20
18	Flow map for support matrix for causal discovery . . . . .	22
19	Support Matrix Explanation . . . . .	23
20	Heat-map for correlation matrix . . . . .	27
21	Numbers of correlation matrix . . . . .	28
22	FGES pairs with alpha change . . . . .	31



23	FGES results in Tetrad . . . . .	32
24	GFCI results in Tetrad . . . . .	33
25	FCI results in Tetrad . . . . .	33
26	PC algorithm pseudo code . . . . .	35
27	PC results in Tetrad . . . . .	36
28	FAS results in Tetrad . . . . .	36
29	Tetrad surface . . . . .	37
30	Two algorithms support the causal relationship . . . . .	38
31	Three algorithms support the causal relationship . . . . .	39
32	Four algorithms support the causal relationship . . . . .	40
33	Five algorithms support the causal relationship . . . . .	41
34	Six algorithms support the causal relationship . . . . .	42
35	Pairs amount declines as algorithm numbers go down . . . . .	43
36	DAG amount vs algorithms support . . . . .	44
37	Nodes amount distribute vs Amount of algorithms that supported . . . . .	45
38	Nodes for five algorithms support the result of Majority Vote . . . . .	48
39	Nodes for six algorithms support the result of Majority Vote . . . . .	48

## PREFACE

Causal discovery is one of the most complex research area comparing to data prediction and data mining. It usually deals with the data which is lack of ground truth in prediction. In social variables, collected data will raise difficulties such as sparsity and inconsistency. In this case, finding a better model for selecting target data is important. While I built the causal discovery models, there are two major challenges. The first technique challenge is the data is not always in Gaussian distribution. For example, the population seems always grows. For time series data, we could arrange the timestamp of the data to make the data Gaussian distributed. In the PC algorithm, the variables are considered as statistic data. Tetrad has a time-lag function which could find the patterns whether the previous year's data have an affection on the following year's data. But for general search, we could still use the Tetrad to train the data as a statistic data.

Thanks to Professor Zadorozhny's guidance from the independent study of social weather system to build the majority vote in causal discovery analysis. The second technique challenge is the causal graph's visualization. Tetrad could deal with one algorithm's causal graph, but couldn't deal with the majority vote's causal graph. With the help of Professor Munro's neural network course's knowledge, I successfully graph the causal relationship using R plot package. Dr. Druzdel also gave me a lot of help in clarifying the difference between the continues data and the time-series data. The causal graph, in this thesis, was described as subgraph for visualization use.

Dr. Glymour gave me lots of help in understanding the latent variables and theory guidance in building the reference model. Special thanks to Clark. Causal discovery is a good direction to select target data, but also relies on the data itself. I am going to explore more about causal discovery in dynamic data in the future.

After one year's work on this project, I learned a lot about how to do research. This train, as far as I could see, is necessary and useful for researchers. From raising a question to finding a way to deal with the research, each progress is unforgettable and memorial. I really appreciate Professor Zadorozhny's efforts and guidance on my project, without his help I couldn't find the right way to deal with the research. I appreciate all the committee members, they are patient and kindness for the time changing of the defense. Special thanks to Clack Glymour, the inventor of PC algorithm who taught me three lessons about the algorithms and how they are functioned. Special thanks to University of Pittsburgh's writing center, Tom and Laura, who helped me editing the thesis's grammar mistakes. Special thanks to my parents. They supported and funded me for an extra half year's tuition. Their financial support helped me pass the durable time. Also, special thanks to my roommates, Fan Yang and Danchen Zhang. Thank you for your support and the company of Fan's cute cat Waffle. Thanks to my neighbour Zhehao Lin took care of my cat Nugget in the busiest time. Thanks to my group members, Hoshang and Swaruba. You helped me a lot in building the system. Thanks to University of Pittsburgh. Three years passed and I really appreciate this school, the opportunity you give to me turns me into a brave woman without fear about research and do what other man could do. Throughout the thorns before road, I will continue my research life without hesitate.

## 1.0 INTRODUCTION

The time-series dataset is widely used in areas such as social research and health-related research. Compared to the 1990s, the open-source datasets are easier to access due to their policy of transparency. Our project based on open-source data from different fields. As we know, weather is influenced by multiple variables such as typhoon, seasons, latitude, etc. Meteorologists, who have already found the patterns through multiple datasets related to weather forecasts, can successfully predict extreme weather, even earthquakes. We take the ideas from the weather forecast to make a social weather forecast. However, social weather is more complex. Social data contain more fields, for example, political news, economic statistics, education, public health, etc. Multiple datasets are stored in different databases; most of them are time-series datasets. Therefore, we tried to build a system collecting the social datasets in a large database, exploring the relationships among the variables without selecting specific target data, see how the data are aggregated and related.

The social weather system we built uses open-source datasets that only focus on the United States. After aggregating the datasets and building the database, we contrasted a front-end system for the database. The database contains datasets from the World Bank<sup>1</sup>, stock market<sup>2</sup>, happiness<sup>3</sup>, terrorism<sup>4</sup> and Internet usage.<sup>5</sup>

Figure 1 is an example of social variables. The arrows represent causes and results between two variables. When the birth expectation increases, the population grows. The growth of population will also influence the customers' consumption, education and employ-

---

<sup>1</sup><https://data.worldbank.org>

<sup>2</sup><https://www.nasdaq.com>

<sup>3</sup><https://worlddatabaseofhappiness.eur.nl>

<sup>4</sup><http://www.start.umd.edu/gtd/>

<sup>5</sup><https://www.internetworldstats.com/unitedstates.htm>

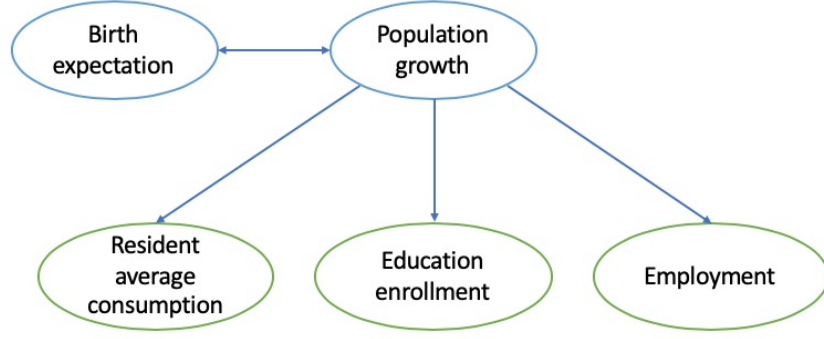


Figure 1: Example of social variables cause and reasons

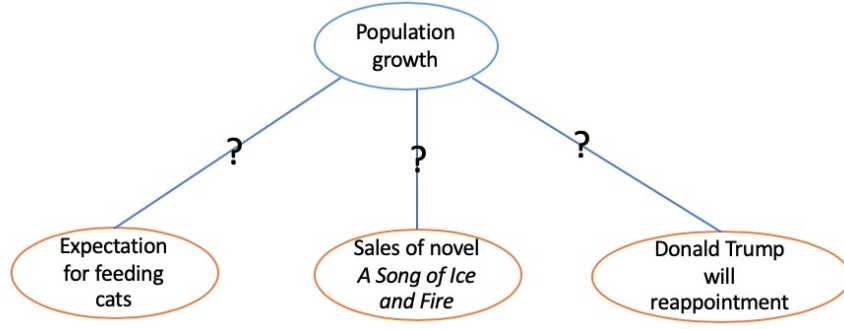


Figure 2: Example of unknowable social variables causes and results

ment. Those patterns, which can be explained as normal senses, have already been analyzed in social models. However, in Figure 2, when we added more datasets in the database, we found it difficult to figure out the patterns of whether population growth will also influence the expectation for feeding cats, sales of a specific book such as *A Song of Ice and Fire* or presidential election results. In this case, we needed to do a causal discovery to predict and cluster the variables which may have relationships. Causal discovery is a branch of pattern mining in data mining; it deals with the similarity between continued variables or discrete variables.

We used a causal discovery software called Tetrad and selected five of the causal dis-

covery search algorithms, which are FCI (Fast Causal Interface), RFCI (Really Fast Casual Interface), FGES (Fast Greedy Equivalence Search) and PC (Peter Spirtes and Clark Glymour) algorithm. Tetrad, which has been under development since the early 1990s, includes a huge variety of tools for causal inference. Tetrad is a special application that contains principled search and predictions of experiment as the users selected the model. It has both exploration surface and a discovery surface for algorithms it offers. For example, when there may be unobserved confounders of measured variables, or models of latent structure, or linear feedback models, Tetrad is good to use. Tetrad is also used in calculating predictions of the effects of interventions or experiments based on a model the user selected.<sup>6</sup> Tetrad is powerful in cleaning datasets, to visualize the causal relationships. After plotting the visualized causal relation map and exporting the pairs in .txt format, we made a support matrix by doing majority vote, using the pre-mentioned five algorithms and the correlation matrix, to identify pairs of data which have a high possibility support of reasons and causes.

While we doing the causal discovery using our time-series data in Tetrad, it automatically consider the data as an independent data. In our research experiment we tried the data with a time-lag as 5 and the amount of the result pairs raises as the time-lag changes from 1 to 5. There are following reasons: 1) Self-loop: The most obvious causal relationship using time-lag is the previous data will cause the following years data. 2) Duplicate in stable data: Some in-normalize distributed data. For example, the forest area seldom changes and if there is a causal saying the carbon dioxide's raise is caused by the decline of the forest area, it is reasonable in common sense. However, the Tetrad concludes the causal relations the surface area's raise is influenced by the forest area. When adding time-lag, it still shows the next few years surface area is the causal of the forest area. 3) Decades may show different patterns: In 1960s to 1980s, the Internet is not as convenient as nowadays and the stock market is also declined in specific period such as the Great Depression.

So in our majority vote model, we focus the data on the following assumption: 1) We assume each variable is independent and not have causal based on year. We just consider the whole variable as one trend line and find the patterns with other variables. 2) For reference model we assume all the variables are acknowledged and have casual connections based on

---

<sup>6</sup><http://www.phil.cmu.edu/tetrad/>

the economic models. 3) We keep the data as the plain data without converting to Gaussian distribution: Tetrad could also dealing and keeping the non-Gaussian distributed data.

We chose the economic growth as a reference data based on the fact that economic growth is already confirmed by the market, industrial expansion and high technology services. According to P. Aghion and P. W. Howitt's economic theories, the economic growth is influenced by several variables.[8] In year 2005, Levine, R. summarizes as follows the existing research on this topic:

Taken as a whole, the bulk of existing research suggests that (1) countries with better functioning banks and markets grow faster; (2) simultaneity bias does not seem to drive these conclusions; (3) better functioning financial systems ease the external financing constraints that impede firm and industrial expansion, suggesting that this is one mechanism through which financial development matters for growth.[9].

Based on the result of their mathematical models, we tried adding more related variables to our databases by training the reference variables our datasets had to determine the performance of the support matrix. Our results show when six algorithms support the results, the reliability of union for datasets and reference reaches 25.92%, and there are seven same pairs with reference (27 pairs) within our database's datasets. The majority vote we use to explore undiscovered causes and reasons among variables that lack ground truth are useful.

## 2.0 LITERATURE REVIEW

Causal discovery is useful in dealing with big data. Data currently expands at a rapid speed. Databases, especially time-series databases, overcome the weaknesses of traditional SQL and non-SQL databases. Causal discovery is a branch of pattern recognition in the data mining area. Patterns is a set of items, subsequences, or substructures that occur frequently together in a dataset. The aim for discovery is to find the inherent regularities (uncovering patterns) in a dataset. However, as the amount of the social news rises, some of the traditional prediction methods show their weaknesses. For example, in traditional economic model, the economists chose the related variables for prediction, but ignored other social data. As the datasets expanded, simply analyzing the economic data limited the exploration of other unknown variables, which may also influence the result. Moreover, more data are required to predict the presidential election, the social activities and the satisfaction of people related to a specific social phenomenon. Therefore, we built the time-series databases by collecting the information from open-source websites.

There are two significant challenges for building the time-series database. The first challenge is the large scale of social data collection and integration. In [16] as Manning, et al mentioned, the data may raise the following questions, resulting in the need for data cleaning: 1) Distributed Data: Social data may be spread over various data sources and organizations. 2) Heterogeneous Data: Social data may be represented in different data formats and in files with different types and structures. 3) Social data may be fragmented, missing values are common. 4) Social data may be aggregated in different ways over various time intervals and space regions. 5) Social data items reported by different data sources may contradict each other resulting in data inconsistencies. 6) Social data sources may have different degrees of reliability. [16]



The second challenge is the traditional pattern mining model, we lack ground truth by combining a variety of datasets. For example, we know the pattern of economic growth based on economic models and we are able to know the pattern for population based on social models, but we lack models for combining the two fields. By adding more fields of data, the accuracy of our reliability declines.

Early in 1997, Cooley, Mobasher and Srivastava provided an overview of tools and techniques for mining the world wide web. They used tools to crawl the data and did data cleaning.[1] In 2000, J Srivastava, R Cooley, M Deshpande, etc provided a more efficient method for pattern discovery. They discribed several steps for pattern discovery: statistical analysis, association rules, clustering the data, classification and use dependency models.[2] In 1997, Meek put forth a conjecture that:

if true, leads to the following and somewhat surprising result: given that the generative distribution has a perfect map in a DAG defined over the observables, then there exists a sparse search space (that is, a space in which each state is connected to a small fraction of the total states) to which we can apply a greedy search algorithm that, in the limit of large number of training cases, identifies the generative structure.[11]

Meek found a method using Greedy Equivilent Search to divide big DAG in relationship graph into small DAGs and enhance the accuracy of causal discovery. Based on the traditional fast greedy search for adjacency matrix, in 2000, Peter Spirtes and Clark Glymour invented an algorithm called the PC algorithm. The algorithm is used on an fMRI experiment to analyze the human brain, and PC algorithm works better than FCI (Fast Causal Inference) in small datasets. The algorithm is sensitive with small datasets.[3] In 2002, David Maxwell proved the Meek Conjecture and improved the Greedy Equivalence Search. [11] In 2008, J. Zhang used FCI to handle causal sufficiency and selection bias. Furthermore, it can be proven that under ideal circumstances it is sound and complete. The FCI algorithms working well on small datasets.[4]

Coumans Vincent(2017) wrote a master thesis, summary and tried PC, FCI and GES on several networks such as Gene and fMRI, but didn't use the algorithm in social databases.[7] As the variables increase, it is hard to control which algorithm behaves the best. That's the reason and motivation we do majority vote. The motivation of our experiment is finding a way to discover the patterns among plenty of social variables, not only in biology. We use

the majority vote to make the support matrix due to these following facts:

- Datasets lack of ground truth to value the behavior of the algorithms.
- A single algorithm may cause too much error.
- Accessed acknowledgement only shows the relation with specific variables in specific area.
- Causal discovery is used in reducing variable amounts while doing prediction and linear regression.

Our aim is to make a support matrix model discovering the causal relationships in a large time-series database. Figure 3 is a flow map for the whole design as the previous description.

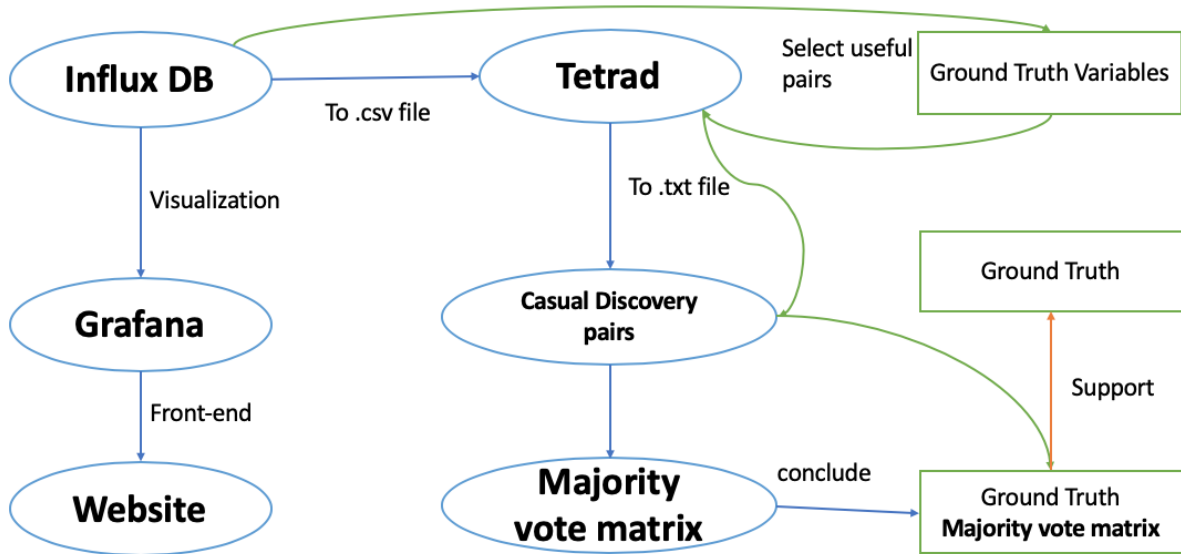


Figure 3: Design flow map

### 3.0 TIME-SERIES DATABASE

Time series data is usually with time lags, it could be yearly, monthly and daily data. The data sometimes are sorted in specific sort method. However, a combination of the time series data needs to do more. Not only collecting the information of the data, we also need to measure a perfect time lag in case of losing information. To provide value, this data requires aggregation and analysis. A time series database is optimized to meet the challenges of handling massive amounts of data from thousands or more devices. For example, Influx DB is one of the best databases we can use for causal discovery. Time-series database also have the following advantages: 1) **Massive scalability and performance:** Effective time series database enables an application to scale easily to support millions time series data points in a continuous flow and perform real-time analysis. 2) **Reduced downtime:** The architecture of a database that is built for time series data ensures that data is always available even in the event of network partitions or hardware failures. 3) **Improved decisions:** Time series database helps an organization make faster and more accurate adjustments important decisions.

In this chapter we explain how the data was collected, how we built the database, how we deal with the front-end and how to visualize the database.

## 3.1 DATA COLLECTION AND DATA CLEANING

### 3.1.1 Data collection

We looked over various open data sources which can be used for this kind of particular task of predicting social outcomes. We finally settled on stocks data, Global Terrorism data, happiness data, Internet data and World Bank data, which includes things such as demographics data, and GDP data. Some data, such as stocks data and happiness data, was collected using web scraping and APIs, while the other data was used in the form of .CSV files. We used python and Jupiter Notebook for web scrapping and further exploration.

### 3.1.2 Data description

- Stock marketing dataset include Russell1000, NASDAQ, SP500 and DJIA.
- The Worldbank dataset contains 51 variables and it is mainly divided into four parts: education (such as admission, education percentage, etc.), energy cost(such as electric use and forest use), population growth(such as birth rate and population) and economy(such as import,export).
- The Internet dataset collects data from year 1990 to year 2017, showing the percentage of the U.S. population people using Internet.
- According to the description of GTD, "The Global Terrorism Database (GTD) is an open-source database including information on terrorist events around the world from 1970 through 2016 (with additional annual updates planned for the future). Unlike many other event databases, the GTD includes systematic data on domestic as well as transnational and international terrorist incidents that have occurred during this time period and now includes more than 170,000 cases. For each GTD incident, information is available on the date and location of the incident, the weapons used and nature of the target, the number of casualties, and—when identifiable—the group or individual responsible."<sup>1</sup>
- According to the description of the happiness dataset, "The happiness datasets contains 12,334 publications in Bibliography of happiness, of which 6,436 report an empirical

---

<sup>1</sup><https://www.start.umd.edu/gtd/about/>

study that is eligible for inclusion in the findings archive. 1167 measures of happiness, mostly single survey questions varying in wording and response scale, 12467 distributional findings in the general public, of which 8703 in 173 nations and 3764 findings in 2455 regions and cities in nations, 2151 studies with findings in 161 specific public's, 15579 correlation findings observed in 2062 studies, excerpted from 1606 publications.”<sup>2</sup>

### 3.1.3 Data cleaning

Data cleaning included steps such as filling null values and normalizing the data. We used python and Jupiter Notebook for this task. We filled 0 for all the numerical null values and filled 'None' for all categorical variables. We also dropped some columns which were used as IDs for some categorical variables as they were providing no significant value.

After filling the null values, we did normalization for all the numerical variables where we converted all the values in terms of percentages. Based on the fact some of the data are different by orders of magnitude, we normalized the dataset by converting the data in terms of percentages since the data came from different sources and we needed to compare the data from these sources.

Then we aggregated the data according to the year of event. We converted the year to make a time stamp and then loaded the data in the Influx DB.

The stock marketing dataset is a frequently changed dataset. Other datasets like the Internet, happiness and World bank datasets, are mainly yearly values. Here we use the average value for the stock marketing to make it a yearly value.

In Figure 4 we can see we have frequently updated data and infrequently updated data. Our method for building the system is we merge all the data together using a same time stamp.

We use the python to load the data in the Influx DB. After loading the datasets could be seen using terminal. Figure 5 is the result in terminal for our database. The data was collected with the same measurement that helps us to do the search queries.

---

<sup>2</sup><https://worlddatabaseofhappiness.eur.nl>

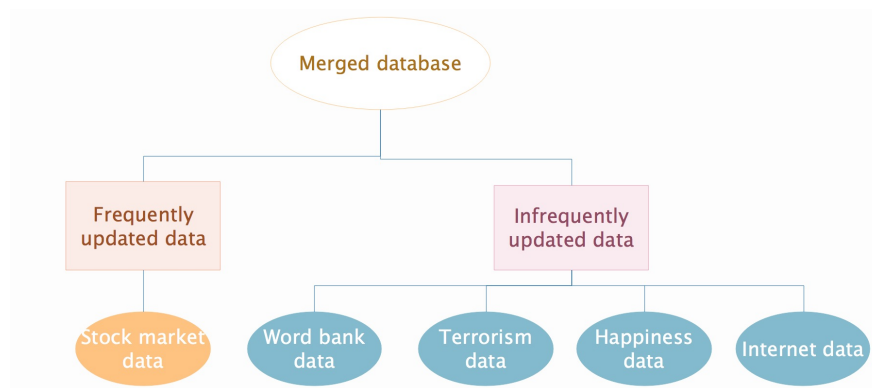


Figure 4: Summary for data aggregation

```

sherry — influx db — 80x24
Last login: Wed Nov 14 02:09:29 on ttys000
-bash: /Users/sherry/.profile: No such file or directory
sherrydeMacBook-Pro:~ sherry$ influx db
Connected to http://localhost:8086 version v1.4.2
InfluxDB shell version: v1.4.2
> show databases
name: databases
name
----
socialweather_norm
_internal
> use databases
ERR: Database databases doesn't exist. Run SHOW DATABASES for a list of existing
databases.
> use databases socialweather_norm
Could not parse database name from "use databases socialweather_norm".
> use socialweather_norm
Using database socialweather_norm
> show measurements
name: measurements
name
----
socialweather
> show socialweather
  
```

Figure 5: The database used a measurement aggregated all datasets

We used Grafana as a visualization method for the database. Grafana is an open source suite, it is used for metric analytics and visualization. It is most commonly used for visualizing time series data for infrastructure and application analytic, and it have some plugin

for research areas such as industrial sensors, global items visualization, weather prediction, and process control. <sup>3</sup> In Figure six, the Grafana showed 49 variables in the system after normalization. Grafana is also used in the calculation of the variables, such as selecting the mean value, plotting a histogram, etc.

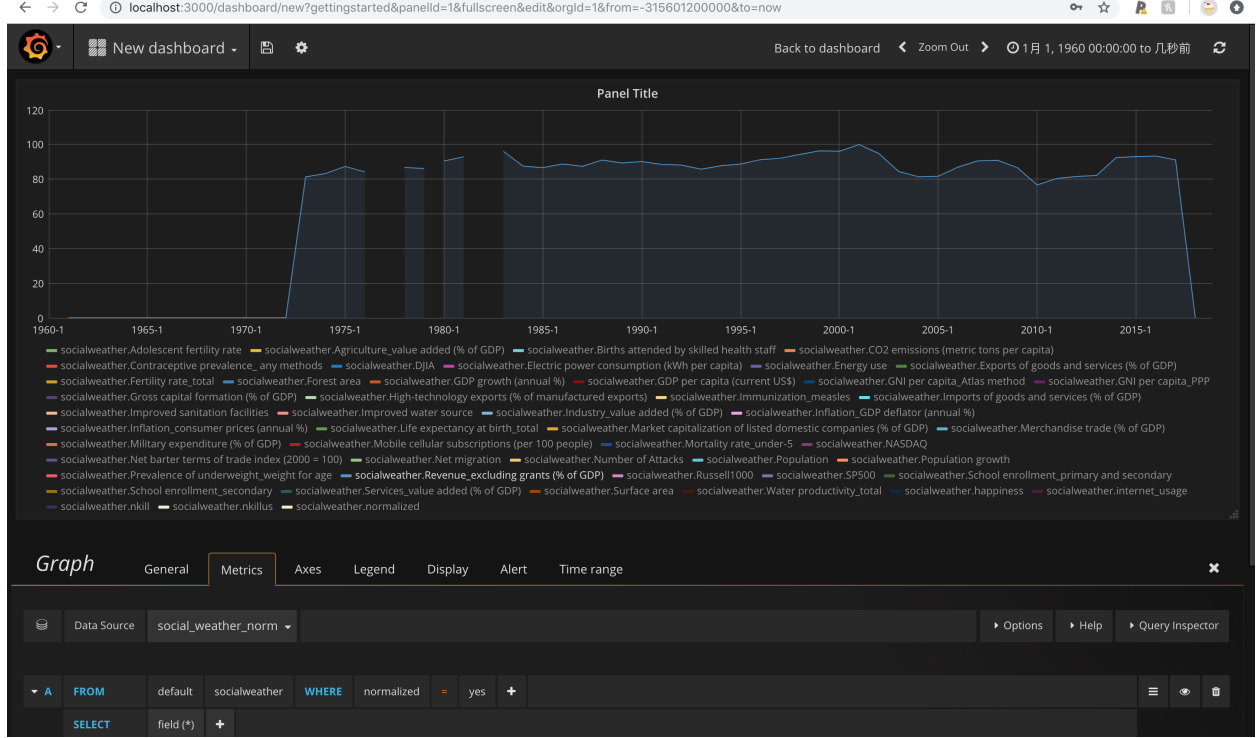


Figure 6: The visualization of Social Weather System in Grafana

### 3.1.4 Front End Design

After we visualized the data in Grafana where we saw several variables on the same plot. Example: You can see n-kill (number of people killed) and nkillus (number of US citizens killed) on the same time series plot.

Also, we converted the longitude and latitude values to geohash values to plot them on the world map on front end. We put these graphs and tables of summary of the incidents on the front end using iframes.

<sup>3</sup><http://docs.grafana.org/v4.3/>

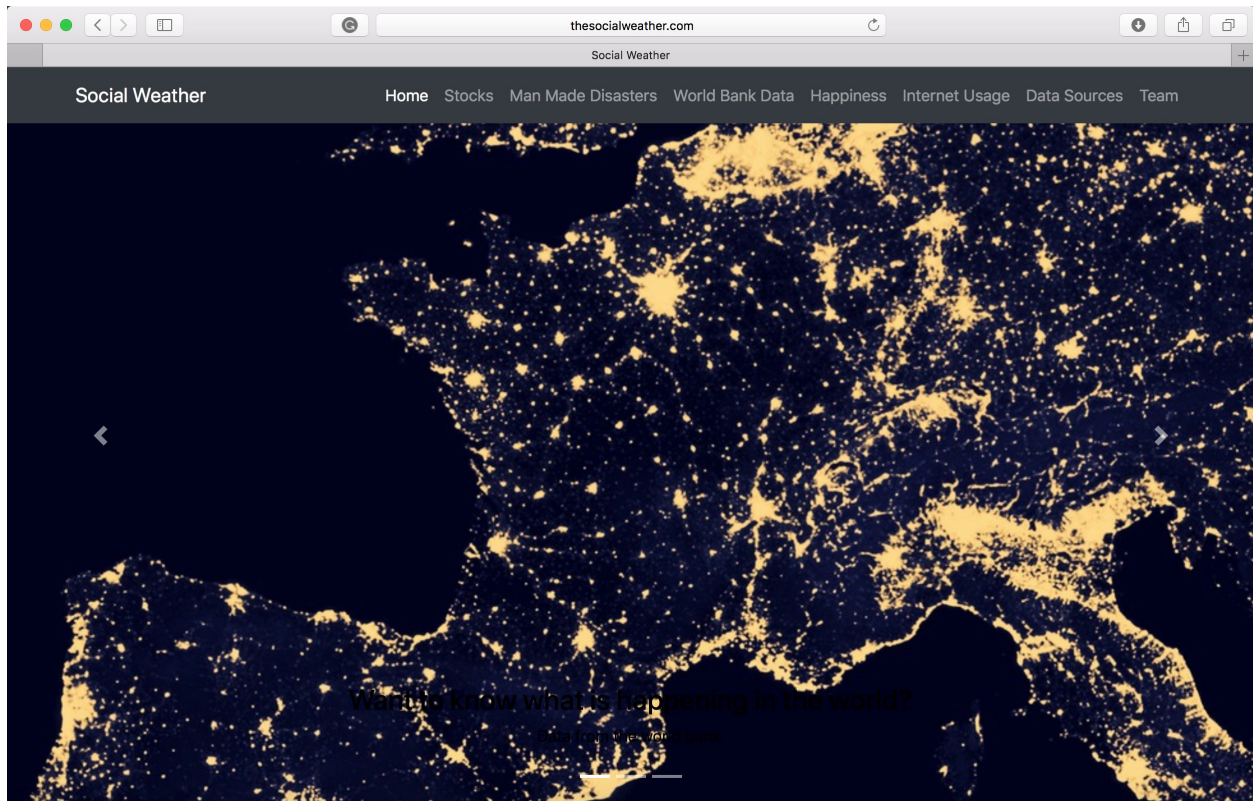


Figure 7: Main Page for Social Weather

Figure 7 is the main page for the social weather station. At the top is the menu bar, all the data from Influx DB are visualized by Grafana. When we want to edit the data, we can click the title of the graph and dial the edit to enter the database.

**3.1.4.1 Front-End Design for Stock market** Stock marketing frequently changes within minutes. We received these data from Quandl, which was first transported inside SQL. After these we created SQL DDL statements to convert the data into dataframes; we turned into .csv files, which we then transported into our time series database InfluxDB. This is made a connector to Grafana, which gave us the better time-series graphs.



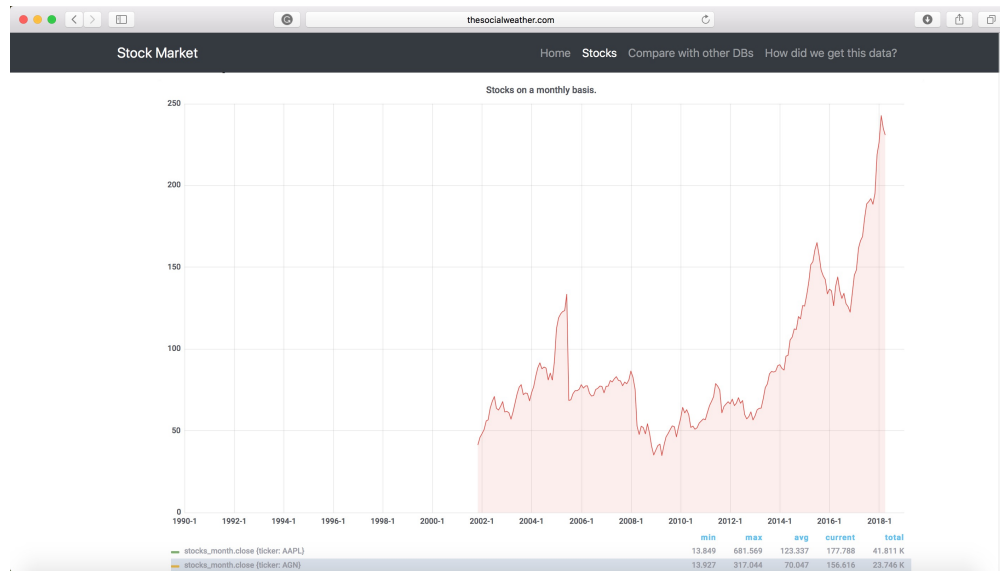
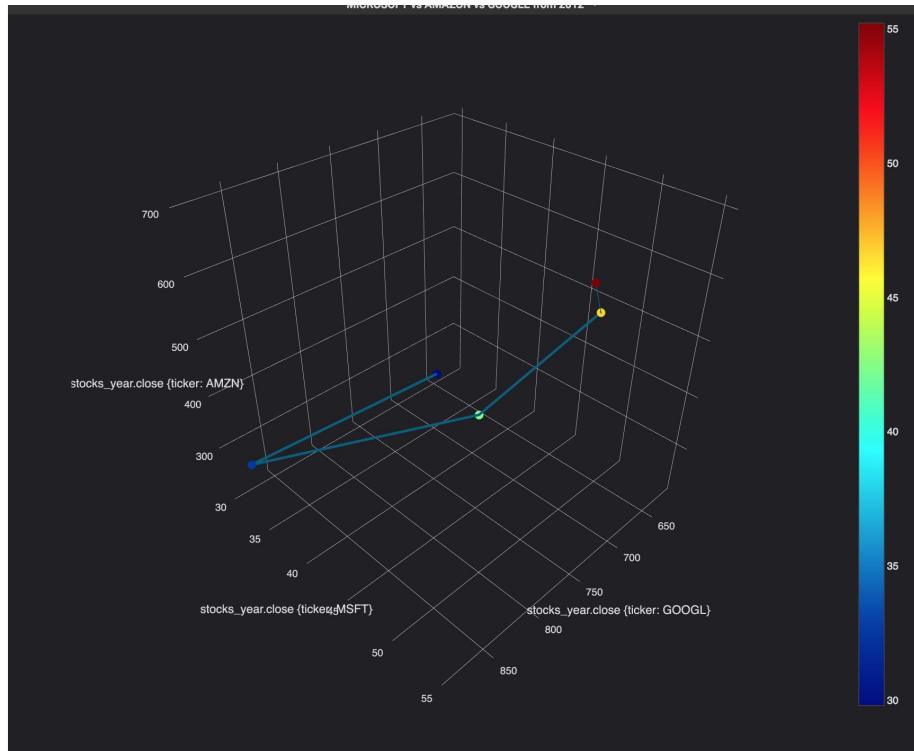


Figure 8: Stock Page - Stocks

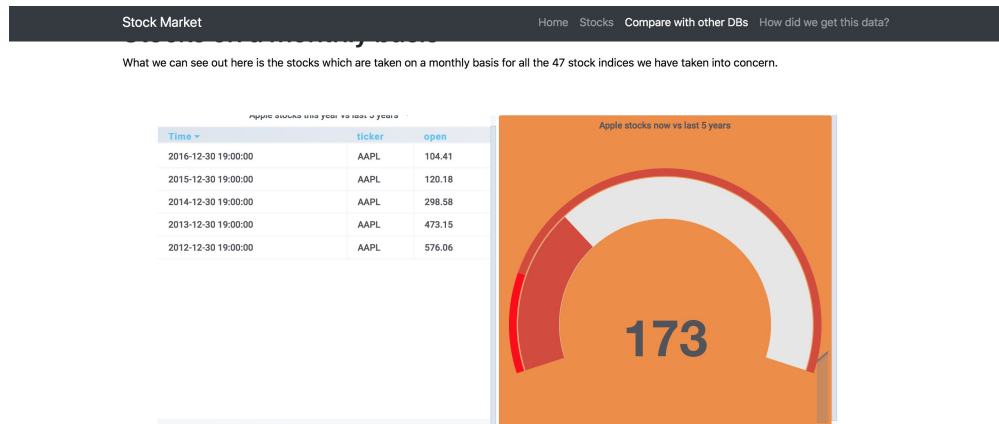
Figure 8 is the stock page for the stock market. This website page gives a summary of the stock database. We have taken the stocks on a monthly basis and checked when they go above a margin at 800. Clicking the button on this page shows query as a table where we have the timestamps along with the values of those stocks with a value greater than 800.

Figure 9 is a 3D graph that illustrates the patterns of these variables. Take Google, Amazon and Microsoft, for example. The color from red to white represents how one company on stock market influences others. The number represents the average value of the stock market per year, for each point in the graph it has a x,y and z value for the three companies. And the time is from year 2012 to year 2017. Figure 10 shows a comparison with two or more databases. Here we can see how the queries work. A "How did we get this data?" at the right menu bar shows how the tables are merged.



Microsoft vs Amazon vs Google from 2012 to 2017(3-D graphs)

Figure 9: 3D graph for Google, Amazon and Microsoft



Apple stocks this year vs last 5 years How good is apple doing?

Figure 10: Stock Marketing by comparison

**3.1.4.2 Front-End Design for the World Bank** After using Grafana to select and compare World Bank variables, we discovered two significant comparisons.

We can see the electric consumption from 1960 to 2015 had a positive relationship with Internet usage. Figure 11 shows a summary of the 41 variables of the World Bank dataset.

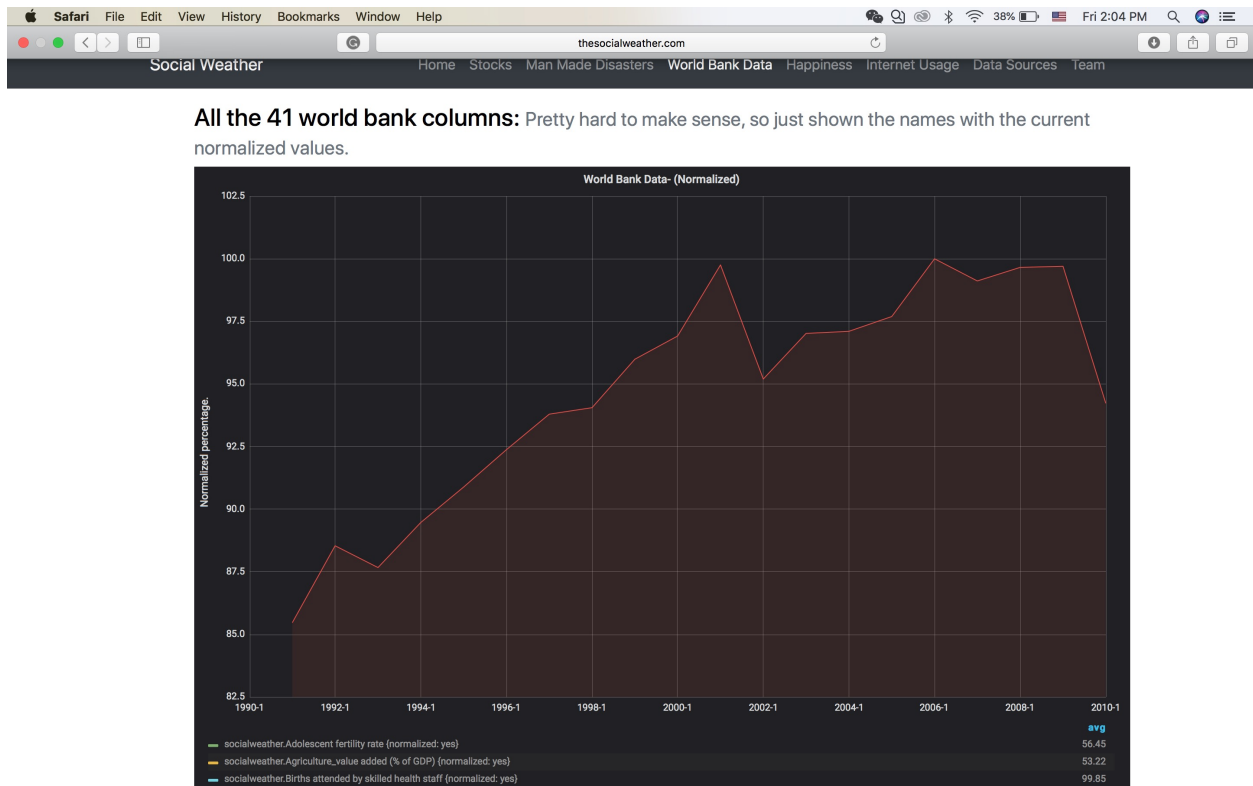


Figure 11: General Data for world Bank

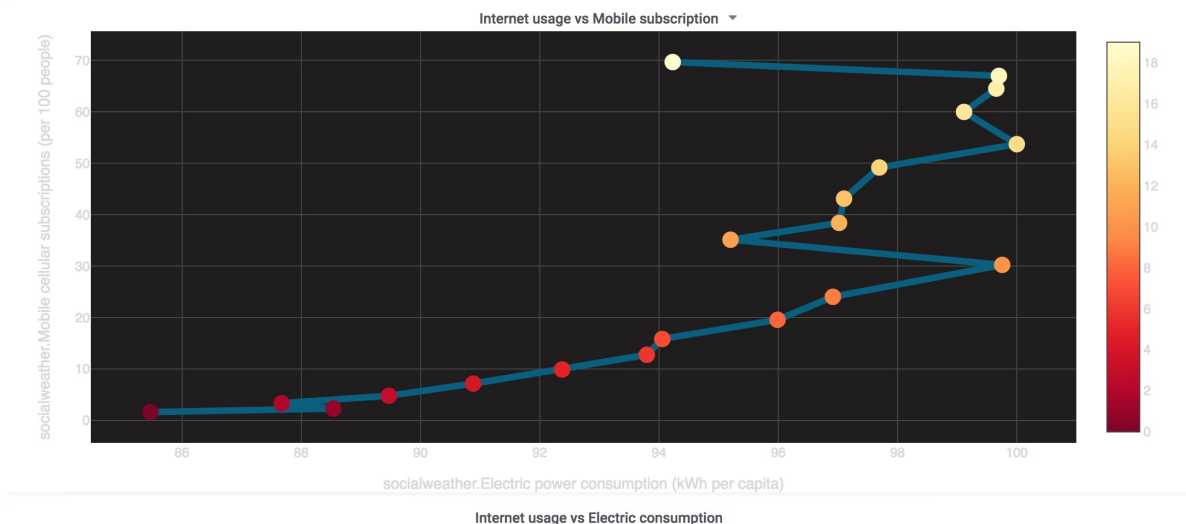


Figure 12: Comparison of Internet Usage vs Mobile Subscription

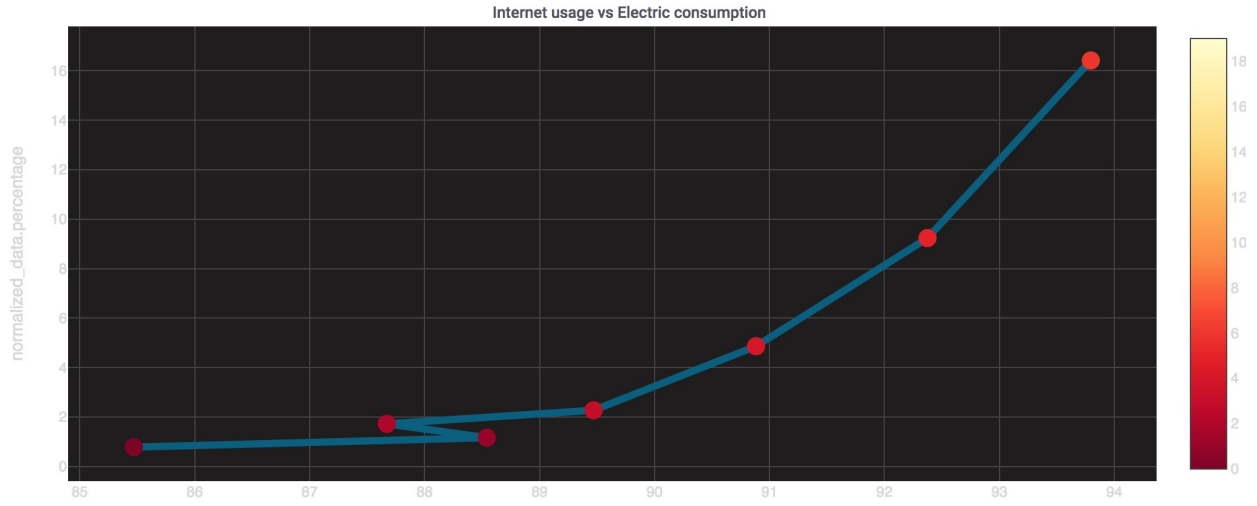


Figure 13: Comparison of Internet Usage vs Electric Consumption

Another interesting comparison in Figure 12 and Figure 13 shows that the growth of Internet usage is not always positively associated with the growth of mobile subscriptions. However, the trend of Internet usage and electric consumption has a positive correlation.

**3.1.4.3 Terrorism visualization** Grafana has a plug-in world-map panel, which could map the Geohashed values of terrorism attacks to the location. Therefore, we used the Python Library PyGeohash, and applied the data frame column of longitude and latitude to receive the Geohashed values.

Figure 14 presents map of the United States. The red points refer to those cities that have experienced terrorist attacks. Clicking the button shows the details of the attack and the consequences.

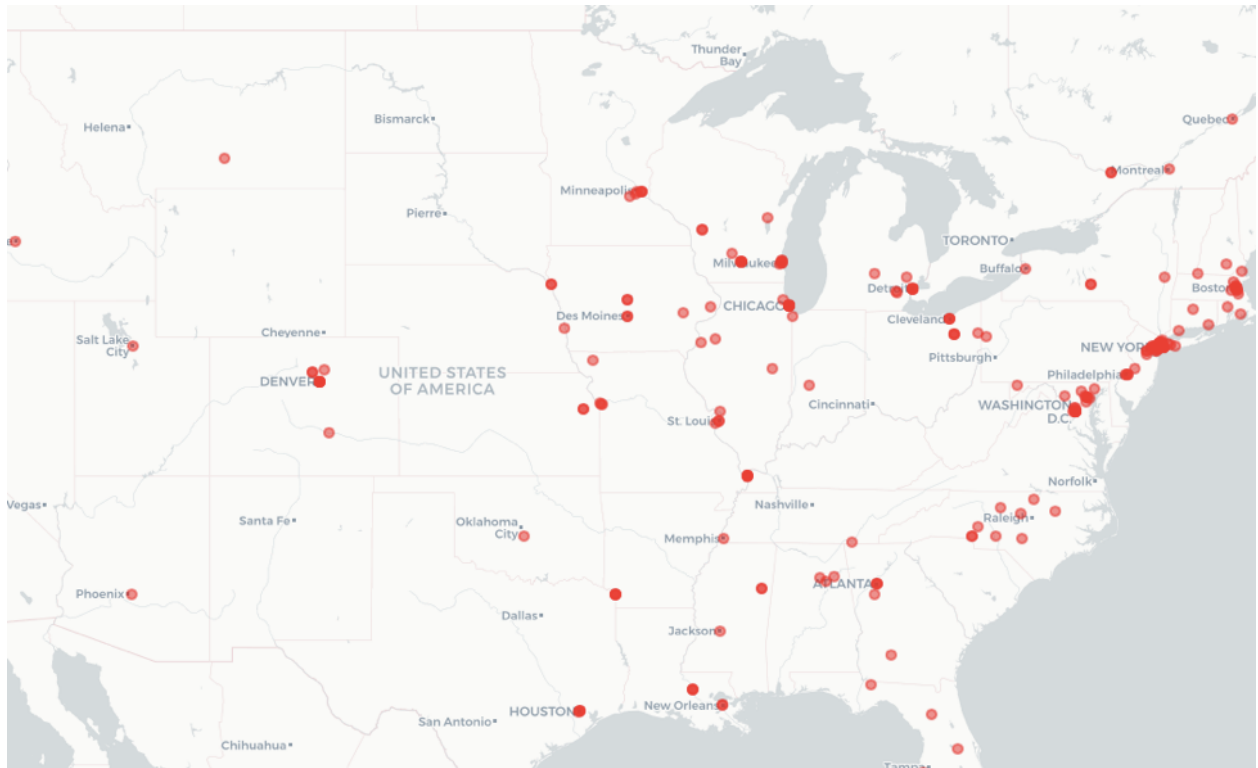


Figure 14: Front End Design for Terrorism Web Page

**3.1.4.4 Front-End Design for Internet Usage** We compared the Internet usage with the stock market. Many people today use the Internet to invest in the stock market. In Figure 15 and Figure 16, the stock market as a "Z" line, which means it initially has a positive correlation with the Internet data negative one later.

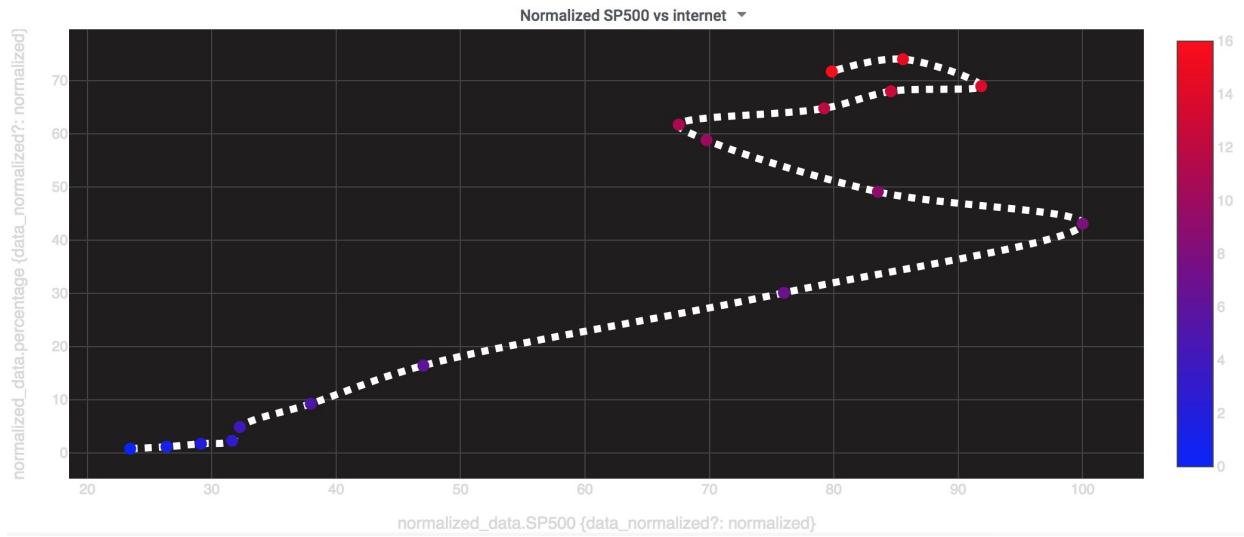


Figure 15: Comparison of Internet with SP500

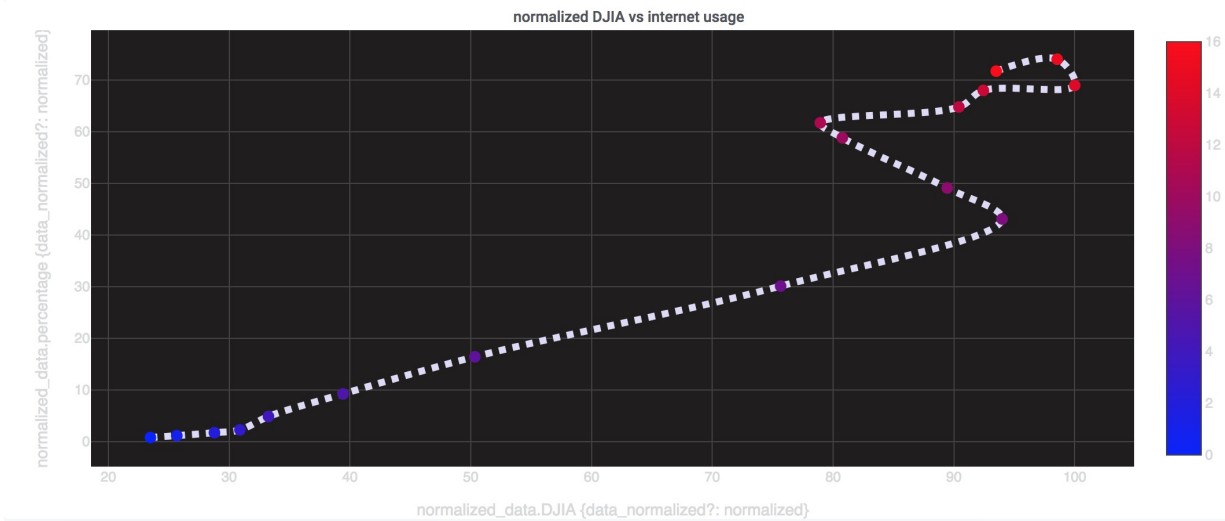


Figure 16: Comparison of Internet with DJIA

### 3.1.4.5 Front-End Design for Happiness

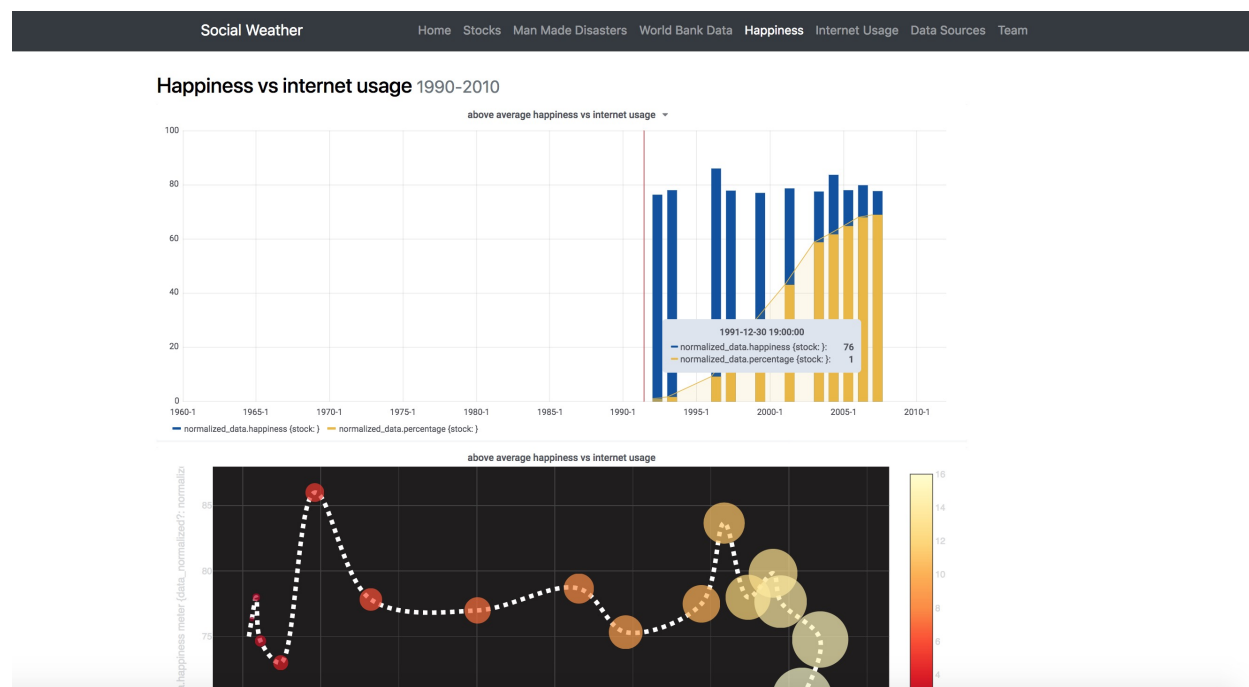


Figure 17: Comparison of happiness and Internet

We used the datasets to compare happiness with Internet usage. In Figure 17, the blue bar represents the average happiness point, and the yellow bar represents the value for Internet usage (total number). The line chart x coordinate symbolized Internet usage, and the y coordinate shows happiness. Happiness and the Internet had a positive correlation in the early years (1960-1970) but a negative correlation in recent years (1990-2017).

## 4.0 CAUSAL DISCOVERY AND SUPPORT MATRIX

After making general pattern comparisons with Grafana, we chose Tetrad to discover the causal relationships among the variables. Tetrad is a desktop Java application that can connect to outside super computing resources if necessary; it creates or simulates data from estimates or tests, and predicts or searches for causal and statistical models. The aim of the program is to provide sophisticated methods in a friendly interface that requires very little statistical knowledge of the user and no programming knowledge. <sup>1</sup>

By using Tetrad we also used five algorithms: Fast Greedy Equivalence Search (FGES), Greedy Fast Causal Inference (GFCI), Fast Causal Inference (FCI), Fast Greedy Equivalence Search (FGES), and PC algorithm and correlation matrix. We transformed the result from Tetrad as a graph matrix. Figure 18 and Figure 19 show how we built up the support matrix using majority voting. The Boyer Moore majority vote algorithm use linear time and constant space to find the majority of a sequence of elements. It is named after Robert S. Boyer and J Strother Moore, who published it in 1981.[10], and is a prototypical example of a streaming algorithm.

In its simplest form, the algorithm discovers a majority element, if there is one:

”An element that occurs repeatedly for more than half of the elements of the input. However, if there is no majority, the algorithm will not detect that fact but will still output one of the elements. A version of the algorithm that makes a second pass through the data can be used to verify that the element found in the first pass is a majority.”<sup>2</sup>

We defined the variable A causing variable B as pair[A, B]. Table 1 lists four pair styles of Tetrad.

---

<sup>1</sup><https://www.ccd.pitt.edu/tools/>

<sup>2</sup>[https://en.wikipedia.org/wiki/Boyer%E2%80%93Moore\\_majority\\_vote\\_algorithm](https://en.wikipedia.org/wiki/Boyer%E2%80%93Moore_majority_vote_algorithm)



Table 1: Definition of Lines for Support Matrix

Edge Types	Present Relationships	Absent Relationship
$A \rightarrow B$	A is a cause of B. It may be a direct or indirect cause that may include other measured variables. Also, there may be an unmeasured confounder of A and B.	B is not a cause of A
$A \leftrightarrow B$	There is an unmeasured confounder (call it L) of A and B. There may be measured variables along the causal pathway from L to A or from L to B.	A is not a cause of B. B is not a cause of A
$A o \rightarrow B$	Either A is a cause of B (i.e. $A \rightarrow B$ ) or there is an unmeasured confounder of A and B (i.e. $A \leftrightarrow B$ ) or both	B is not a cause of A
$A o - o B$	Exactly one of the following holds: 1. A is a cause of B 2. B is a cause of A 3. there is an unmeasured confounder of A and B 4. both a and c 5. both b and c	

Table 1 defines the edge type. We give  $[1,1]$  value to  $A o - o B$  and give  $[1,0]$  value to  $A \rightarrow B$ . 1 represents the cause in our model. Since the type  $A \leftrightarrow B$  infers that the pairs have a latent reason L that causes these two variables, we defined these kind of pairs as  $[0,0]$ . While we explored the time series datasets with one algorithm, we learned that the behavior of a single algorithm is not as good as predicted due to the interaction of fields of datasets. In this case, we did majority vote to analyze the potential relationships of our datasets.

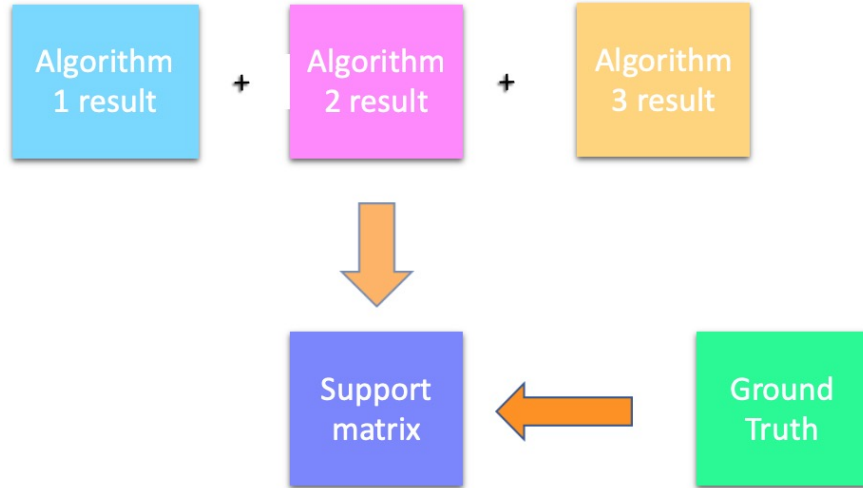
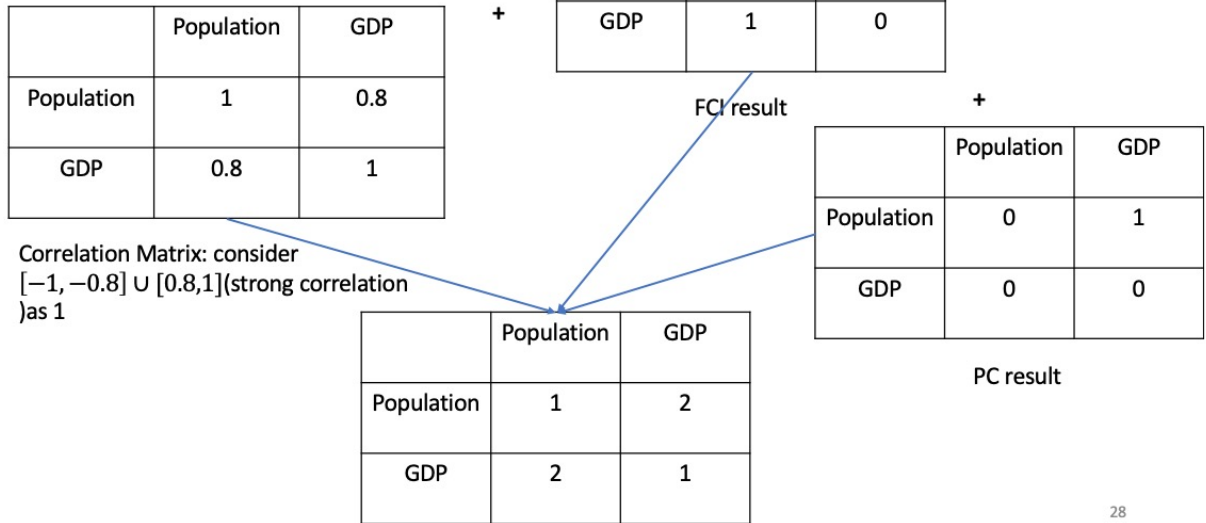


Figure 18: Flow map for support matrix for causal discovery

# Support Matrix



28

Figure 19: Support Matrix Explanation

Figure 19 shows the support matrix for causal discovery value for FCI and PC. By adding the value of the matrix, we defined the support matrix as a majority vote for the six algorithms.

## 5.0 MATHEMATICAL FOUNDATION

### 5.0.1 DAG

We defined  $v$  as our datasets, and the real causal relationship will result in directed acyclic graph (DAG) which is transitive, irreflexive and anti-symmetric. For example, three variables  $X$ ,  $Y$ , and  $Z$ , when  $X \rightarrow Y$  and  $Y \rightarrow Z$ , we can define the relationship  $X \rightarrow Y \rightarrow Z$  as a DAG.

### 5.0.2 d-separate

D-separation is a relation between three disjoint sets of vertices in a directed graph. The basic idea involves checking whether a set of vertices  $Z$  blocks all connections of a certain type between  $X$  and  $Y$  in a graph  $G$ . If so, then  $X$  and  $Y$  are d-separated by  $Z$  in  $G$ . By choosing d-separation to connect DAGs to probability distributions, we assume that in all of the distributions  $P$  a DAG  $G$  can represent. If sets of vertices  $X$  and  $Y$  are d-separated by a set  $Z$  in the DAG  $G$ , then  $X$  and  $Y$  are conditional independent of  $Z$  in  $P$ . We called the relationship  $X \rightarrow Y \rightarrow Z$  as  $X \perp\!\!\!\perp Z | Y$  by d-separate, which means that  $X$  and  $Z$  are conditionally independent.[\[13\]](#)

### 5.0.3 Causal Markov Condition

Holland in 1986 defined the causal Markov Condition is :

A variable  $X$  is independent of every other variable (except  $X$ 's effects) conditional on all of its direct causes.[\[13\]](#)

According to the Causal Markov Condition we defined the  $X \perp\!\!\!\perp Z|Y$  as when  $Y$  grows the  $Z$  will grow no matter the growth and decline of  $X$ . The DAG should follow the rules as Vincent summarized:

1. A DAG  $G$  and a probability distribution  $p$  satisfy the Causal Markov Assumption if each variable is conditionally independent of its non-descendants given its parents. In this case we say that  $p$  is generated by  $G$ .
2. A DAG  $G$  and a probability distribution  $p$  that satisfy the Causal Markov Assumption satisfy the Faithfulness Assumption if each conditional independence statement valid for  $p$  is implied by the Causal Markov Assumption. In this case we say that  $G$  and  $p$  are faithful to one another.

Given a DAG  $G$  that is faithful to a probability distribution  $p$ , one can read conditional independence statements from the DAG via d-separation.<sup>[7]</sup>

#### 5.0.4 Markov Equivalence

Given DAGs  $G$  and  $H$ , then  $G$  and  $H$  are said to be Markov equivalent if for every probability distribution  $p$  we have that  $p$  and  $G$  are faithful to one another if and only if  $p$  and  $H$  are faithful to one another.

#### 5.0.5 Skeleton

Let  $G$  be a graph on nodes  $V$ . Then the skeleton of  $G$  is the undirected graph on  $V$  such that for every pair of nodes  $X, Y \in V$  one has that  $X$  and  $Y$  are adjacent in  $G$  if there is an edge between  $X$  and  $Y$  in the skeleton of  $G$ .

#### 5.0.6 CPDAG

Let  $M$  be a Markov Equivalence class of DAGs. Then the pattern representing  $M$  is a graph whose skeleton is the same as the skeleton of all the DAGs in that Markov equivalence class and such that for every pair of adjacent nodes  $X, Y$  we have that the edge between  $X$  and  $Y$  is a directed edge  $X \rightarrow Y$  if for every DAG in the equivalence class that edge is oriented as  $X \rightarrow Y$ ; otherwise it is an undirected edge.

### 5.0.7 Bayesian network

A Bayesian network is a probabilistic graphical model (a type of statistical model) that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG).[\[17\]](#) In our model we have five algorithms based on the Bayesian network. There is a weakness in using the Bayesian network for causal discovery: while we are using the Bayesian network to analyze the datasets, assuming we have two variables  $a$  and  $b$ , the result is  $a \leftrightarrow b$ . That is the reason we tried five other algorithms in addition to that using only Bayesian network. A causal network is a Bayesian network with the requirement that rather than the relationships be causal, which means that our experiment results rely on some variables that have high possibility to have causal and result patterns. To deal with this problem, these algorithms do a "cutting off" to the connected pairs.

The additional semantics of causal networks specify that if a node  $X$  is actively caused to be in a given state  $x$  (an action written as  $\text{do}(X = x)$ ), then the probability density function changes to that of the network obtained by cutting the links from the parents of  $X$  to  $X$ , and setting  $X$  to the caused value  $x$ .[\[12\]](#)

Using these semantics, the impact of external interventions from data obtained prior to intervention can be predicted.

## 6.0 CAUSAL DISCOVERY ALGORITHMS

### 6.0.1 Correlation

We removed the time stamp and calculated the correlation value. Figure 20 shows a heat map correlation matrix. The number from 0 to 48 represents the 49 variables. The range of correlation matrix's value is  $[-1,1]$ . The color changes from light green to blue. The darker the color is, the higher positive relevant the variables are. We chose the variables whose absolute value which was larger than 0.8 regarding them as good behavior in causal discovery.

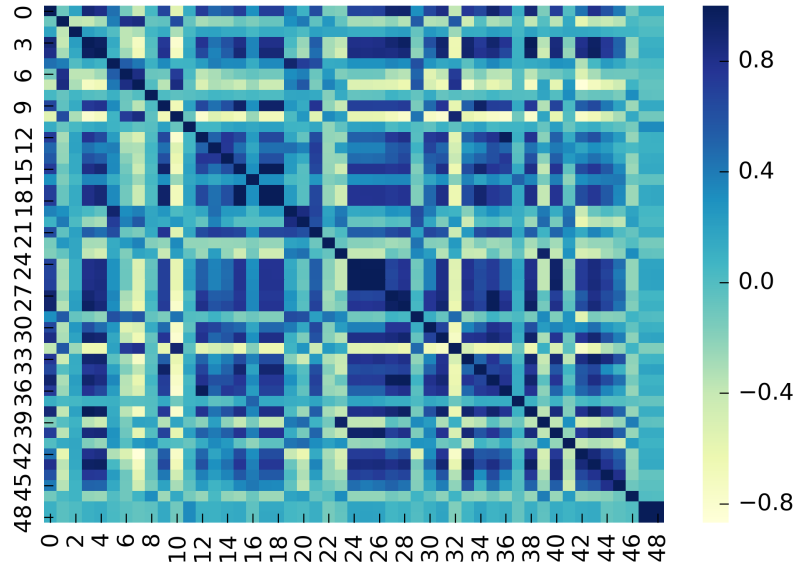


Figure 20: Heat-map for correlation matrix

Figure 21 is the summary for the correlation matrix. The red line indicates how many correlation pairs' absolute values exit as the value changed from  $[0,1]$ . We can see in the map when we selected 0.8 to train the support matrix since around 10% of the pairs represent the data that has a stronger connection. The other five algorithm models we used for our datasets have a similar amount of pairs. Therefore 0.8 is enough as a symbol for building the support matrix.

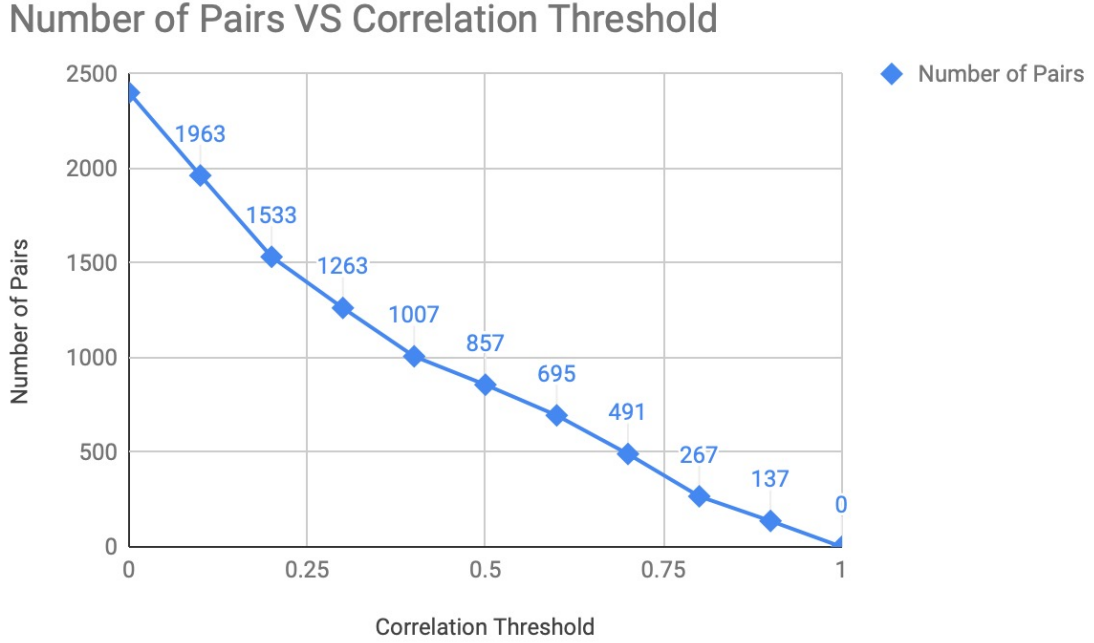


Figure 21: Numbers of correlation matrix

### 6.0.2 Fast Greedy Equivalence Search

Fast Greedy Equivalence Search for continuous data (FGESc) is an algorithm that takes as input a dataset of continuous variables, greedily searches over selected causal Bayesian network (CBN) structures(models), and outputs the most probable model it finds. The model FGESc returns serve as a data-supported hypothesis about causal relationships that exist among the variables in the dataset. The model is intended to help scientists form hypotheses

and guide the design of controlled experiments to investigate these hypotheses.<sup>1</sup> FGESc originated from 1997, when Meek developed an algorithm called the Greedy Equivalence Search (GES). This algorithm was based on an assumption with two directed acyclic graphs (DAGs), graph 1 and graph 2. GES is a Bayesian algorithm that heuristically searches the space of CBNs and returns the model with the highest Bayesian score it finds. In particular, GES starts its search with the empty graph. It then performs a forward stepping search in which edges are added between nodes in order to increase the Bayesian score. This process continues until no single edge addition increases the score. Finally, GES performs a backward stepping search that removes edges until no single edge removal can increase the score.[11] In order to explain this part, we have a pseudo code to see how GES deals with CPDAGs. Listings 1 and 2 illustrate how GES adds and removes the lines until the the score does not further increase.

```
function add(data.csv, CPDAG graph1)
{Consider the set S of graph obtainable from graph1 by adding one edge;
Let x be a DAG from S with the largest s(data.csv, x);
if S(data.csv,graph1)<S(data.csv,x) then
x is graph2
else
graph1 is graph2
end
}
return graph2
```

Listing 1: GES adds edges as long as the score increases

---

<sup>1</sup>[https://www.ccd.pitt.edu/wiki/index.php/Fast\\_Greedy\\_Equivalence\\_Search\\_\(FGES\)](https://www.ccd.pitt.edu/wiki/index.php/Fast_Greedy_Equivalence_Search_(FGES))



```

function remove(data.csv, CPDAG graph1)
{Consider the set S of graph obtainable from graph1 by removing one edge;
Let x be a DAG from S with the largest s(data.csv, x);
if S(data.csv,graph1)<S(data.csv,x) then
x is graph2
else
graph1 is graph2
end
}
return graph2

```

Listing 2: GES removes edges until the score does not change

Listing 3 shows how the GES works after the balance.

```

function GES(data.csv, CPDAG graph1){
graph 1 = empty graph;
while S(data.csv, add(data.csv, graph1) >s(data.csv,graph1) do
graph1 = add(data.csv,graph1)
end
while S(data.csv, remove(data.csv, graph1) >s(data.csv,graph1) do
graph1 = add(data.csv,graph1)
end
return graph1

```

Listing 3: GES adds edges until the score does not change

FGESc uses the Bayesian Information Criterion (BIC) [15] score models, the score models approximate the marginal likelihood of the data given a graph structure  $M$ :  $P(\text{data}|M)$ . More

Table 2: Table for rules of data

columns represent variables, rows represent samples,variable in a sample is continuous
the data and variable names are separated by a delimiter,in order and unique
no missing values in the table.
no linear dependencies in the data
no variables that have zero variance

precisely, the score models approximate the natural logarithm of the marginal likelihood. The data require the following rules: In Table 2<sup>2</sup>; our data perfectly match all the requirements. After we tested the alpha value in the function, we found these patterns:

### Pairs for FGES VS Penalty Discount

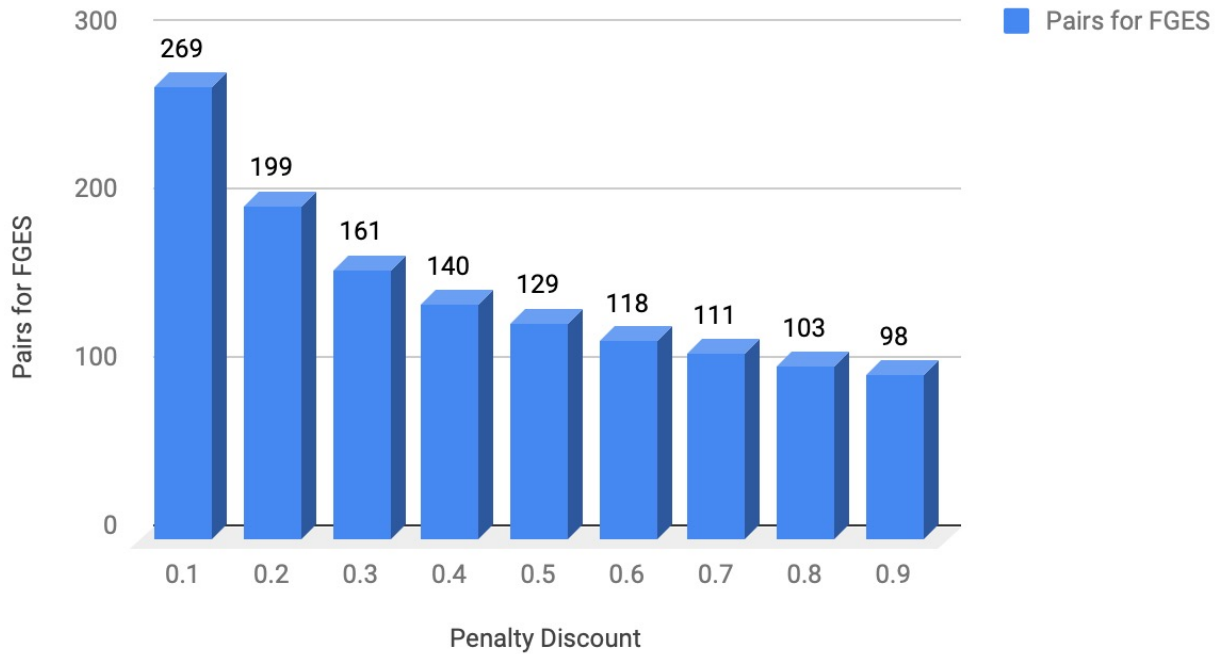


Figure 22: FGES pairs with alpha change

<sup>2</sup>[https://www.ccd.pitt.edu/wiki/index.php/Fast\\_Greedy\\_Equivalence\\_Search\\_\(FGES\)\\_Algorithm\\_for\\_Continuous\\_Variables](https://www.ccd.pitt.edu/wiki/index.php/Fast_Greedy_Equivalence_Search_(FGES)_Algorithm_for_Continuous_Variables)

Figure 22 show the number of pairs of nodes with alpha value from  $[0,1]$ ; if the penalty discount reaches 0.1, then the existing pairs will approach near 10% among all the pairs. In this case, we choose  $\text{penalty} = 0.1$  as our test value. Figure 23 is the result in Tetrad using the FGES method. The arrows show how this algorithm works.

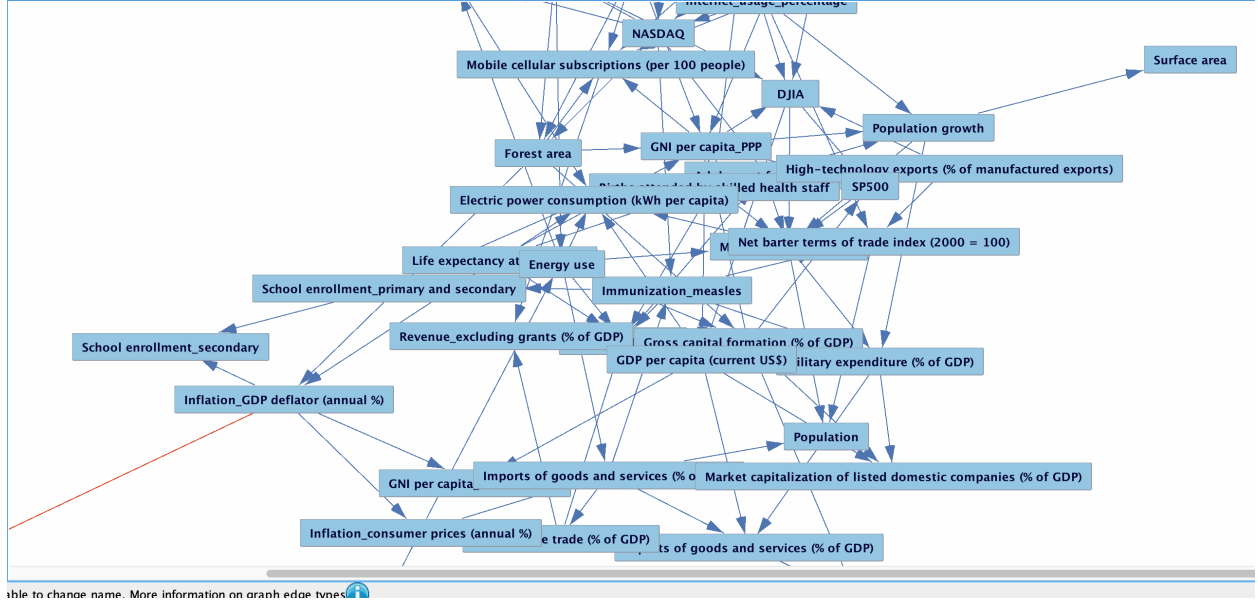


Figure 23: FGES results in Tetrad

### 6.0.3 Greedy Fast Causal Inference (GFCI) and Fast Causal Inference (FCI)

GFCI, which combines the FGES algorithm and the FCI algorithm improves the accuracy and efficiency of FCI. GFCI is an algorithm whose input is a dataset of discrete variables with two phases. The first phase greedily searches over selected causal Bayesian network (CBN) structures (models); it next outputs the highest scoring model it finds under the assumption that there are no unmeasured cofounders and selection bias. This output is then input into a slight modification of the Fast Causal Inference (FCI) algorithm, which post-processes the output to produce a representation of a set of models that may include unmeasured cofounders. The model that GFCI returns serves as a data-supported hypothesis about causal relationships that exist among the variables in the dataset. This type of

to investigate these hypotheses.[\[14\]](#)

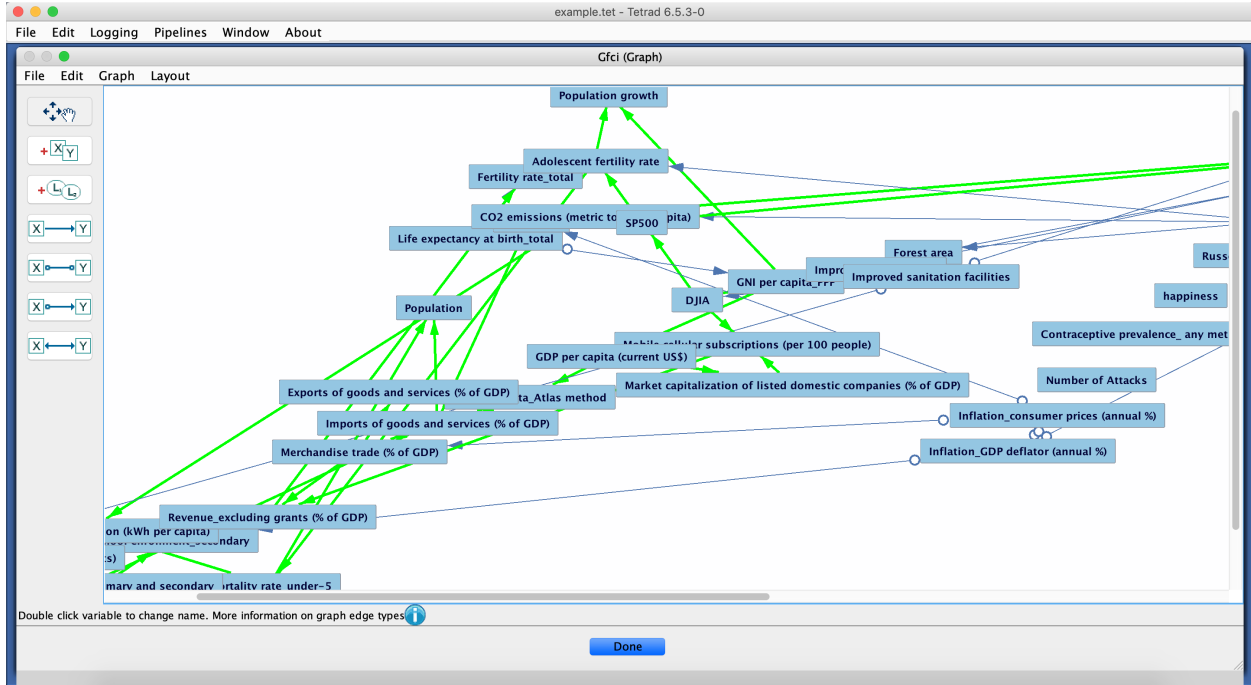


Figure 24: GFCI results in Tetrad

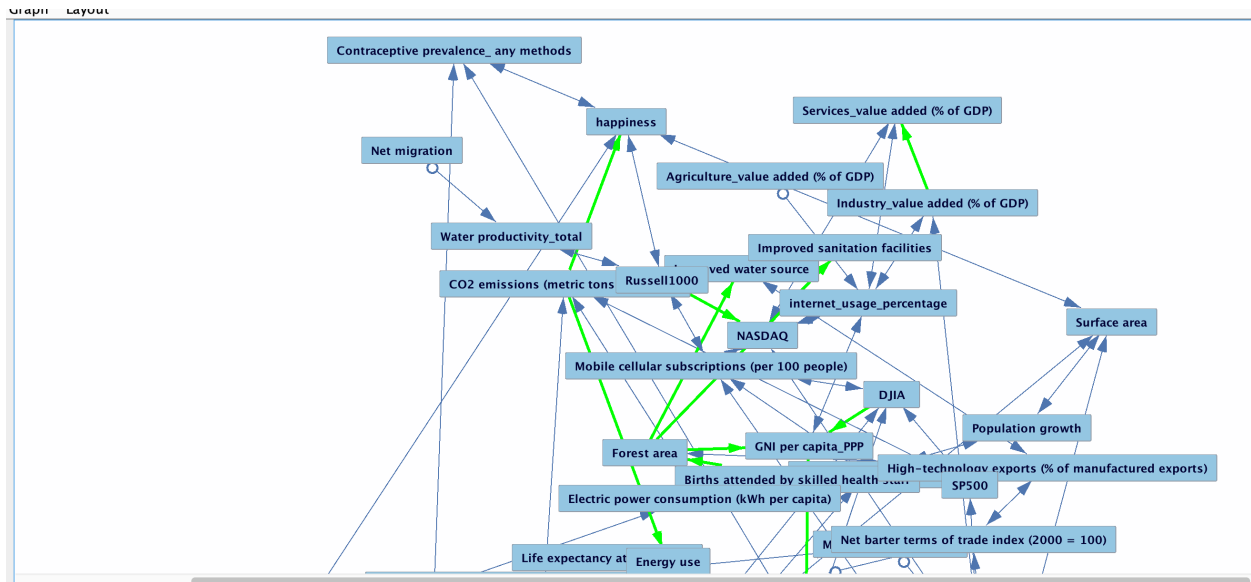


Figure 25: FCI results in Tetrad

FCI algorithms are considered as constraint based algorithms. Different from FAS, constraint-based algorithms first identify several constraints, which the underlying causal structure should satisfy. These constraints are then used to derive the causal structure. The fore mentioned constraints might, for instance, consist of conditional independence statements. Figure 24 and Figure 25 show results from Tetrad for FCI and GFCI.

According to Colombo, et al.(2012),under the faithfulness assumption, a Markov equivalence class of DAGs with latent and selection variables can be learned from conditional independence information among the observed variables alone using the Fast Causal Inference (FCI) algorithm, which is a modification of the PC algorithm.[6]

#### 6.0.4 Fast Adjacency Search(FAS) and PC Algorithm

FAS is based on a search using an adjacency matrix. This is the adjacency search of the PC algorithm, included here for times when only the adjacency search is needed, as when one is subsequently going to orient variables pairwise.

PC algorithm [3] is a pattern search; it assumes that the underlying causal structure of the input data is acyclic, and that no two variables are caused by the same latent variable. Here we talked about modified PC dealing with time continuous datasets. The algorithm assumed that the causal relation between any two variables is linear, and all the variables are normal distribution. In Tetrad, the PC algorithm will sometimes output double-headed edges. In the large sample limit, double-headed edges in the output indicate that the adjacent variables have an unrecorded common cause, but PC tends to produce false positive double headed edges on small samples. In this case, we made a  $[0,0]$  pairs for our data.

The PC algorithm is correct whenever decision procedures for independence and conditional independence are available. The procedure conducts a sequence of independence and conditional independence tests, and efficiently builds a pattern from the results of those tests. According to Sprites, et al. [3], The algorithm-modified PC consists of two phases: the skeleton phase and the orientation phase. In the skeleton phase, conditional independence statements are derived from the data.[3] The resulting skeleton satisfies the following property: for every pair of vertices X and Y, X and Y are adjacent for every subset C of

vertical, not including  $X$  and  $Y$ ; given  $X$  is not conditionally independent from  $Y$ .

In order to explain this, we use the pseudo code from Vincent(2017) to find out how the PC algorithm works. Figure 26 shows how the codes worked.[7]

---

```

input : A dataset  $D$  over variable set  $V$  consisting of  $n$  variables
output: A PAG on  $V$ 
1  $C \leftarrow$  Complete graph over  $V$  where every edge is oriented as o-o;
2 for  $k$  from 0 to  $n$  do
3   for  $X, Y$  distinct adjacent nodes in  $C$ , such that the number of nodes adjacent to  $X$  is greater than
    $k$  do
4     for Every set  $W \subset V \setminus \{Y\}$  of nodes adjacent to  $X$ , with  $\#(W)=k$  do
5       if The conditional independence test on  $(X, Y, W)$  returns TRUE then
6         Erase the edge between  $X$  and  $Y$ ;
7         Remember  $W$ ;
8       end
9     end
10  end
11 end
12 for Unshielded triples  $X * - * Y * - * Z$  in  $C$  do
13   if  $Y$  is not in one of the sets encountered that renders  $X$  and  $Z$  independent then
14     Orient  $X * - * Y * - * Z$  as  $X * \rightarrow Y \leftarrow * Z$ 
15   end
16 end
17 Apply further orientation rules until no more edges can be oriented.
18 return  $C$ 

```

---

Figure 26: PC algorithm pseudo code

The tests have an alpha value for rejecting the null hypothesis, which is always a hypothesis of independence or conditional independence. For continuous variables, the PC uses tests of zero correlation or zero partial correlation for independence, or conditional independence respectively. The tests require an alpha value for rejecting the null hypothesis, which can be adjusted by the user. Usually in small datasets, we chose 0.1-0.2 as the alpha; we tried both of these value and did 0.1 since the pairs are around 10% of the total amount of adjacency matrix pairs. Figure 27 and Figure 28 give the results for Tetrad.

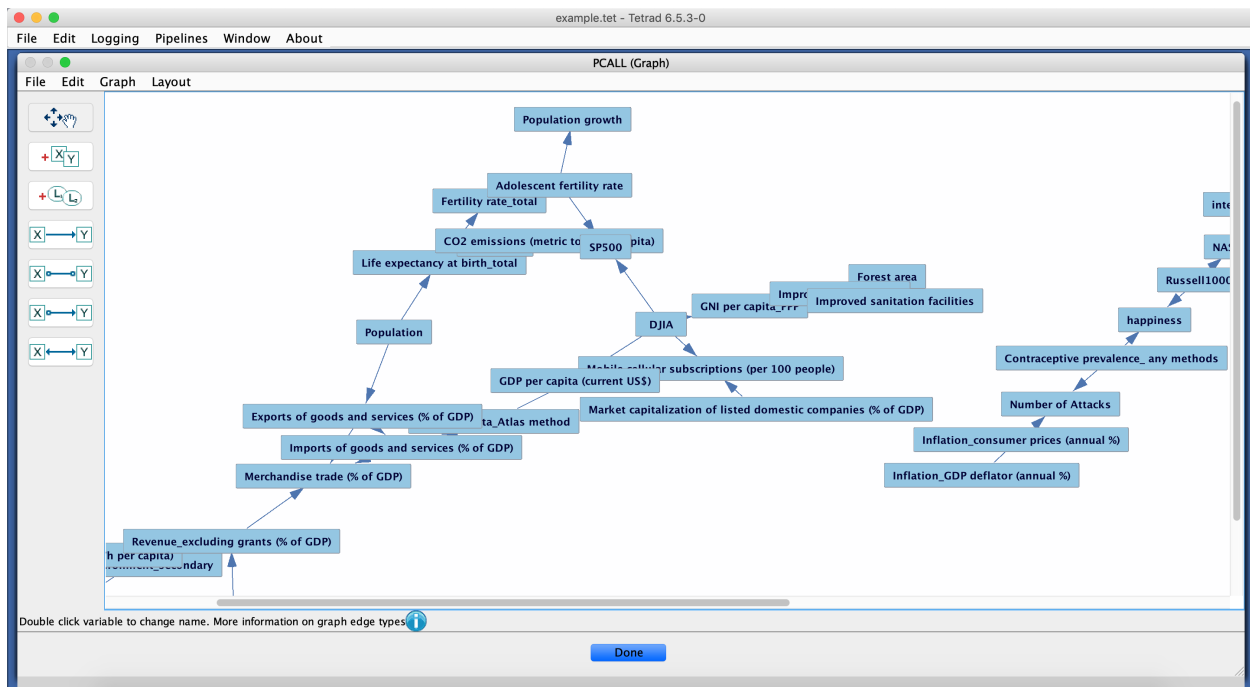


Figure 27: PC results in Tetrad

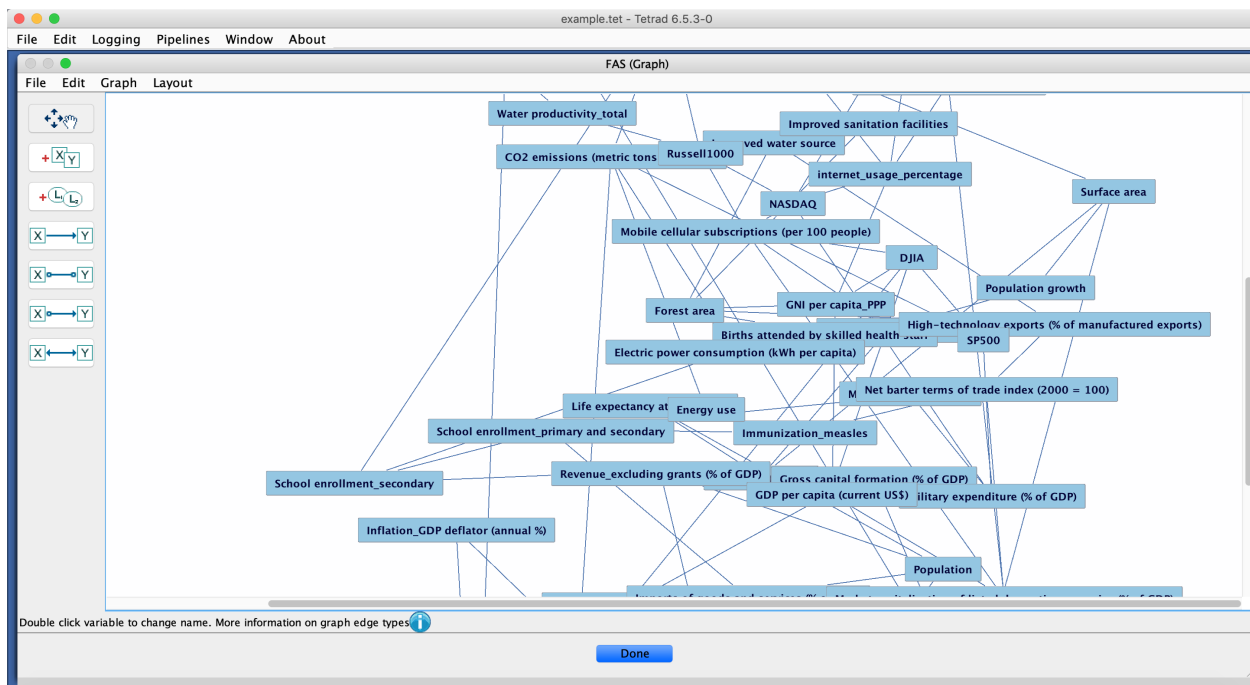


Figure 28: FAS results in Tetrad

**6.0.4.1 Majority vote** We used majority vote for building up the support matrix. We did a sum for the results for social weather database's datasets.

Tetrad plots pairs of discovered sets. It also provides functions for time lag plot and normalizing. Figure 29 shows the flow map of how Tetrad works. After we collected results, we converted the visualized map into an adjacency matrix. After adding the adjacency values, we got a support matrix.

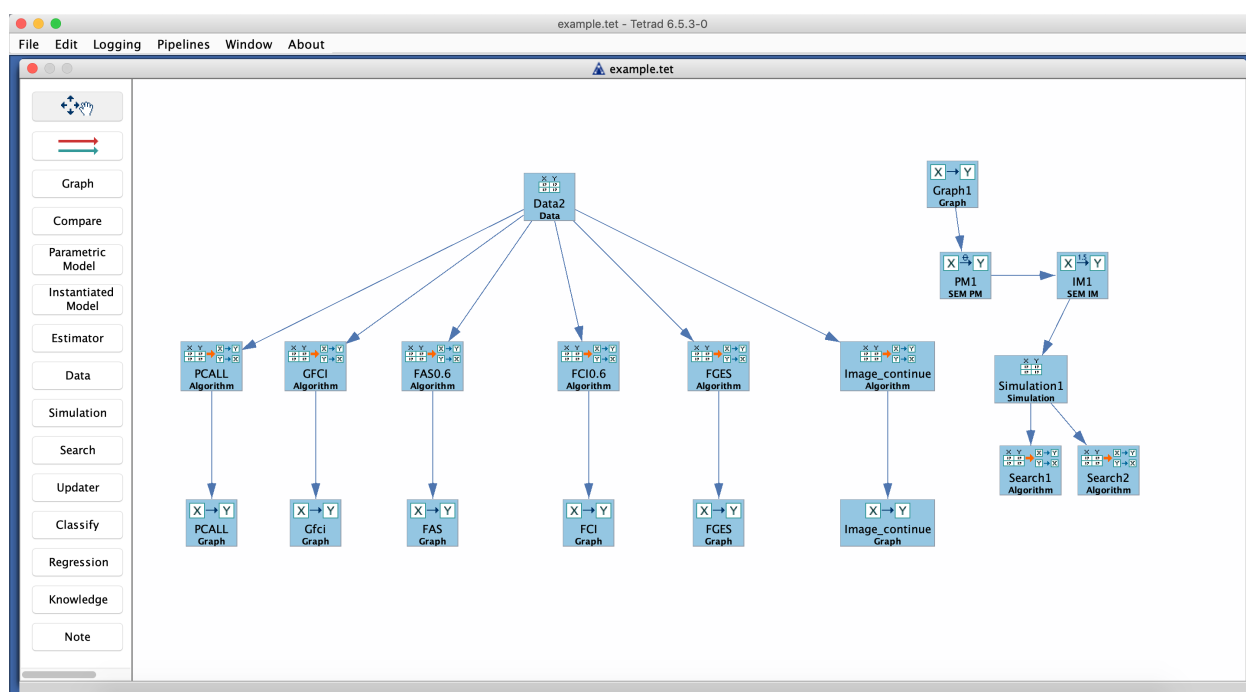


Figure 29: Tetrad surface

We used R to plot the directed acyclic graph (DAG). Figure 30 to Figure 34 show from two algorithms to six algorithms how many algorithms support the inference for causal discovery. Each directed acyclic graphs has a direction, and the value of the definition is either 0 or 1.



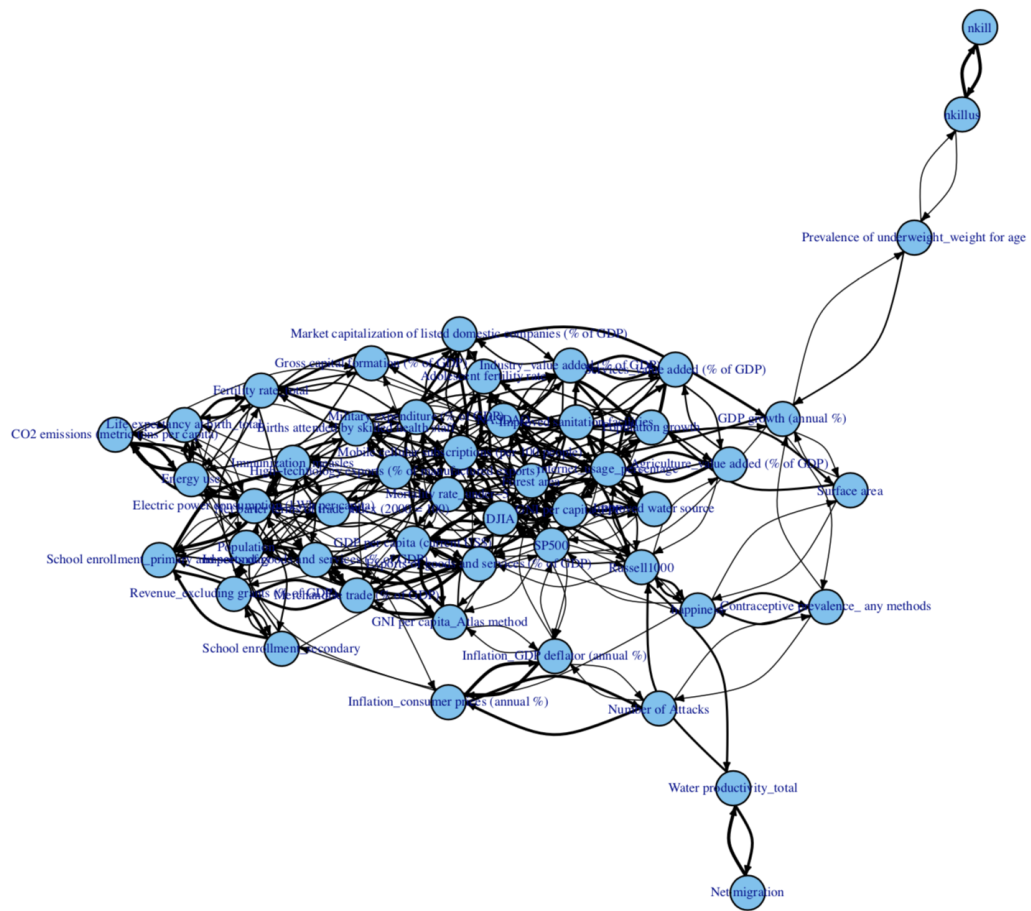


Figure 30: Two algorithms support the causal relationship



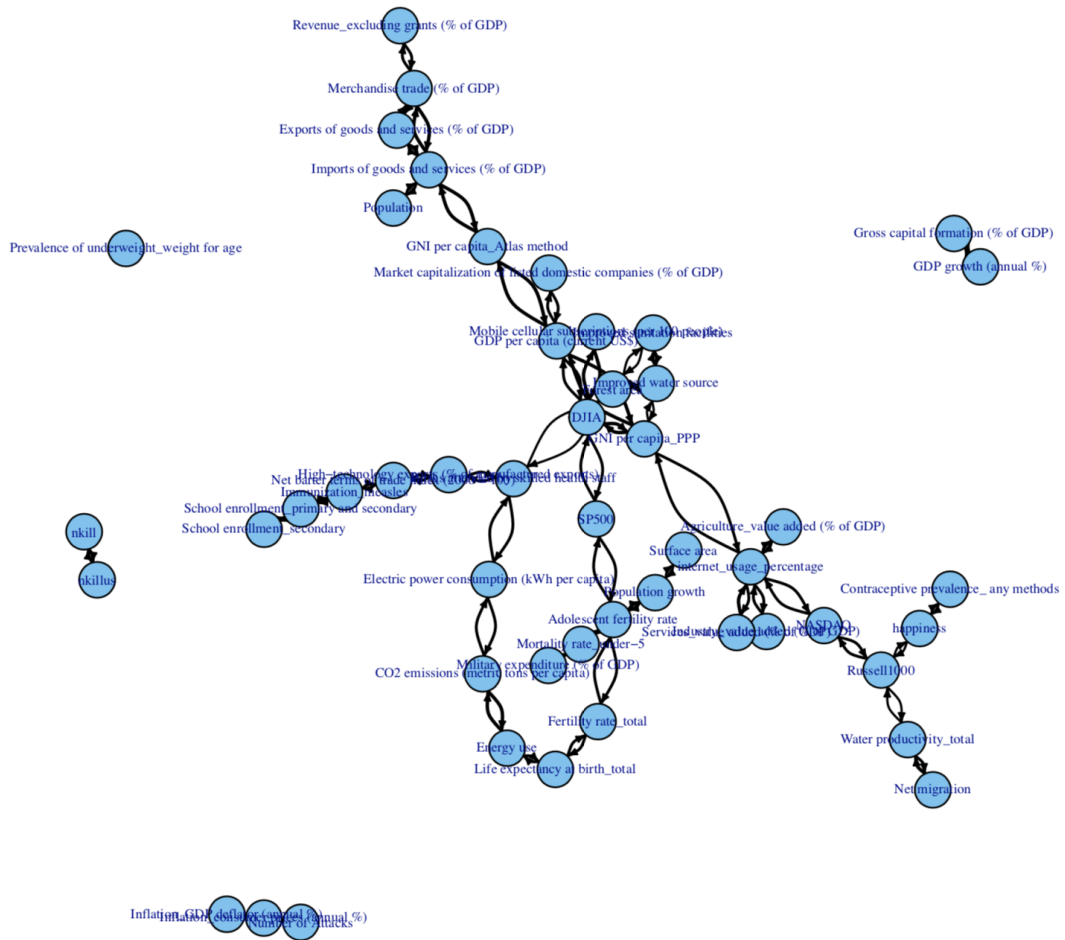


Figure 32: Four algorithms support the causal relationship

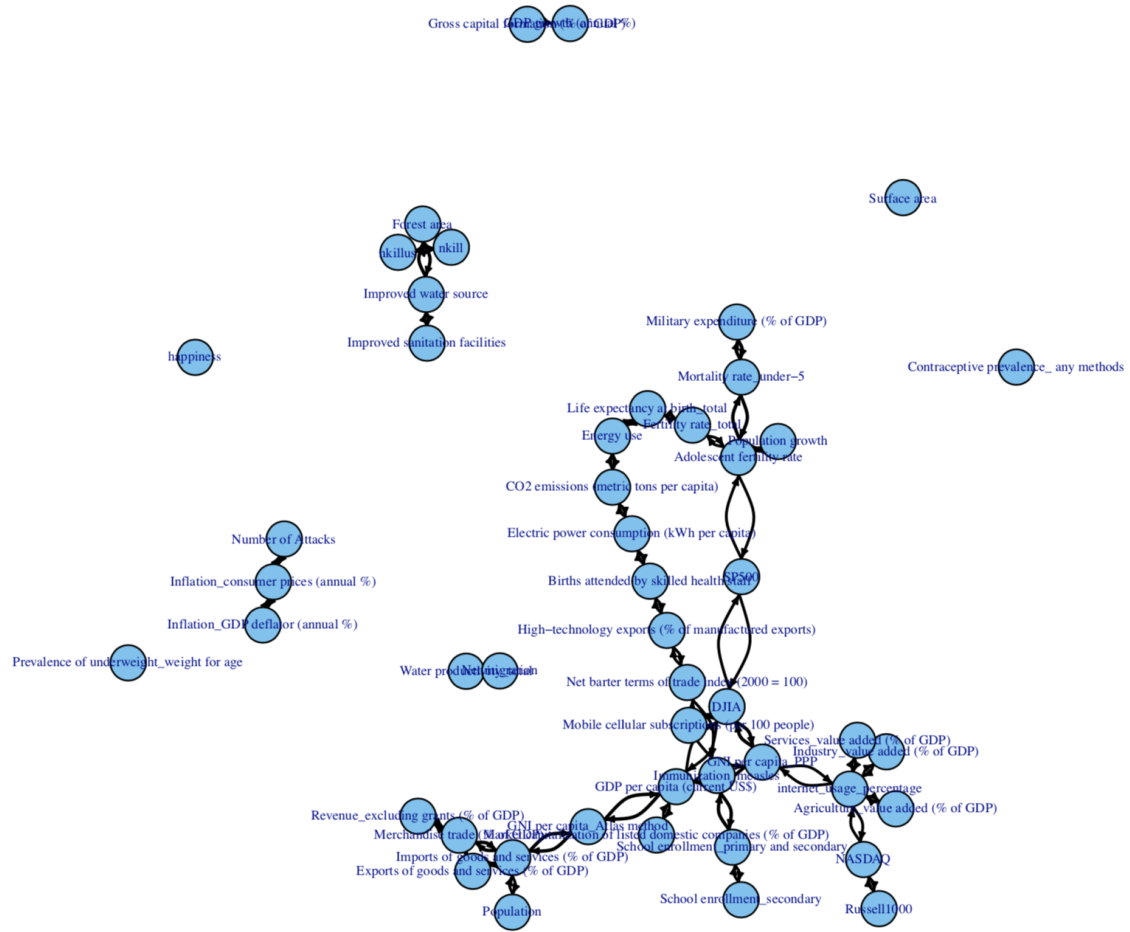


Figure 33: Five algorithms support the causal relationship

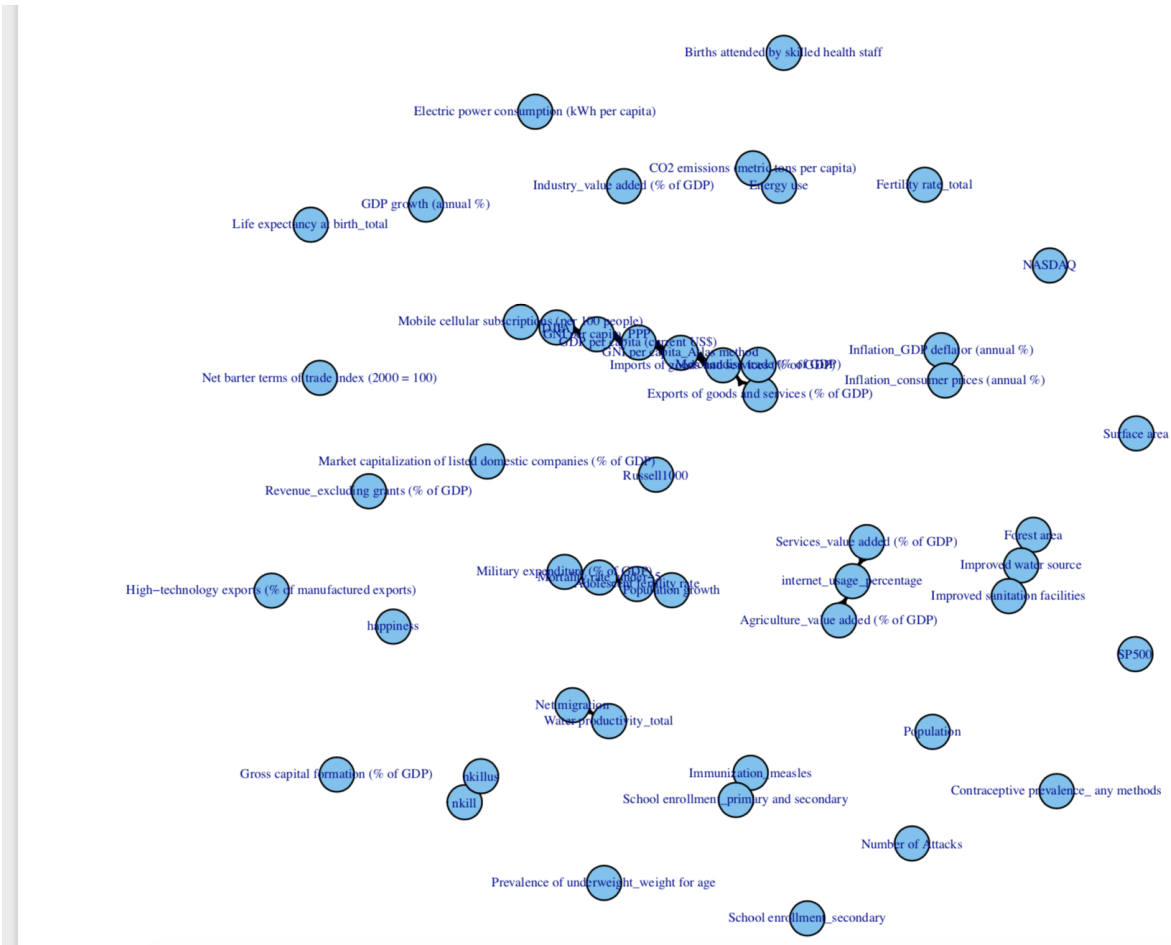


Figure 34: Six algorithms support the causal relationship

Table 3: Pairs of nodes by algorithm support

Numbers of Algorithms support discovery	pairs
>0	979
>1	312
>2	203
>3	98
>4	83
>5	36
Amount	2401

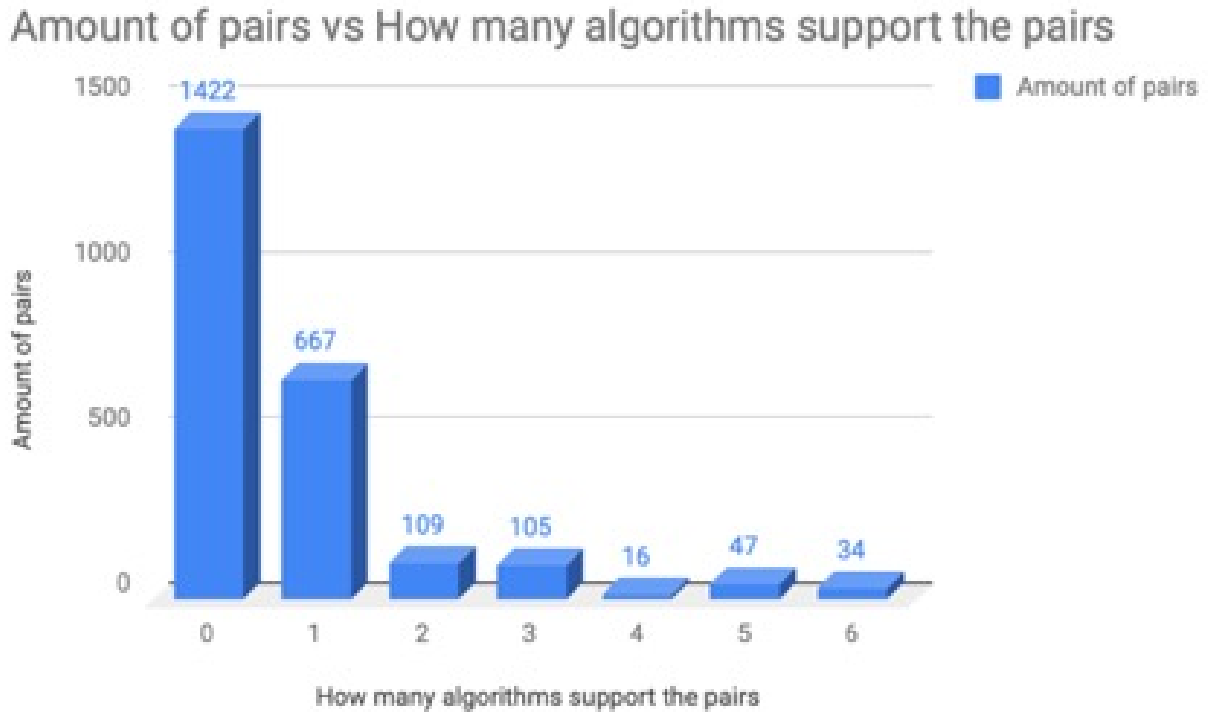


Figure 35: Pairs amount declines as algorithm numbers go down

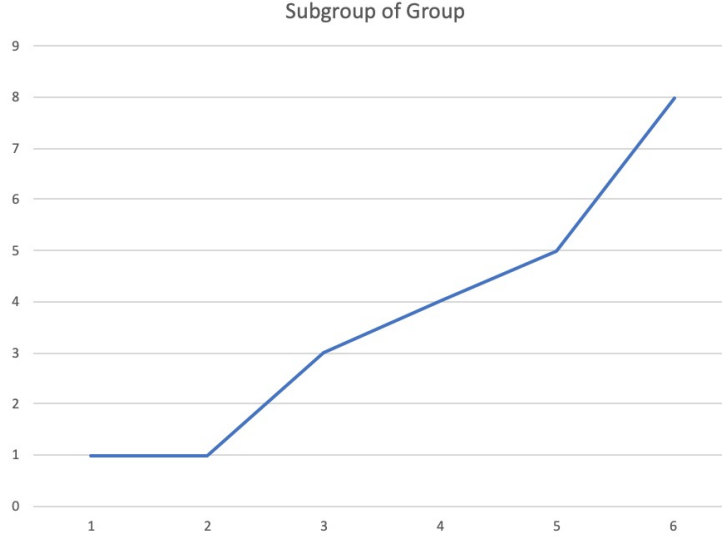


Figure 36: DAG amount vs algorithms support

After the definition of pairs we calculated how many pairs in the matrix. Table 3 shows the pair value. As we counted, we dropped the value of pairs equal to 1, since in the correlation matrix the nodes will have a 1 value with itself. When the majority vote reaches to three algorithms support the inference, and the pairs reaches about 10% of the total pairs amount. Figure 35 shows how the pairs decline as the number of the algorithms rises. Figure 36 shows how many DAGs by algorithm support from [0,6]. The result is 1 when algorithm is 1 and 2 due to the correlation matrix's nodes' self-loop and FAS's adjacency matrix when the dataset is small. As previously mentioned, when there are 3 algorithms or more, the DAG grows.

Table 4: Pairs of nodes by algorithm support

Amount of algorithms that supported majority votes	Nodes amount distributed
>0	49
>1	49
>2	43,3,2
>3	41,2,3,2
>4	32,3,2,3,2
>5	23,4,8,2,3,2,2,2,2

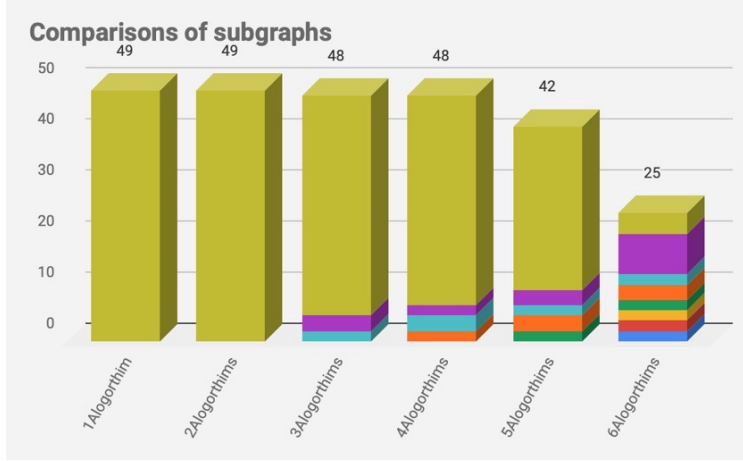


Figure 37: Nodes amount distribute vs Amount of algorithms that supported

Table 4 and Figure 37 show how many nodes are in one DAG graph. The minimum loop, which contains two nodes. This result removed the self-circle of the nodes.

## 6.1 EVALUATE

### 6.1.1 Reference data and Model Behavior

The main idea of our model is to find the reason and cause within the datasets. However, in our dataset, some of the data values are gradually increasing, and the increase follows a linear regression. In this case, we need to rebuild the data with star schema using the same id year. In order to find the ground truth, we considered the star schema based on our previous experiment. We used the target data in Table 5 and removed the stable dataset and the data that need text mining. By adding more data from Fred<sup>3</sup>, we built a new dataset. In case the data cleaning will influence the result of the reference dataset, we explored the dataset to remove the data that are not Gaussian distributed.

Continuous datasets should include two rules to avoid noisy data. The first three rules are: 1)The series have no zero meaning. 2) The variance changes over time. 3)The values

<sup>3</sup><https://fred.stlouisfed.org/>



correlate with lag values.

Based on the selected rule, we changed the way to clean the data, replaced the missing value with the mean average, and tried time lag 1 with the data. We added from Fred. According to Aghion and Howitt's economic theories, the economic growth is influenced by several variables.[8] In 2005, Levine summarized the existing research on this topic:

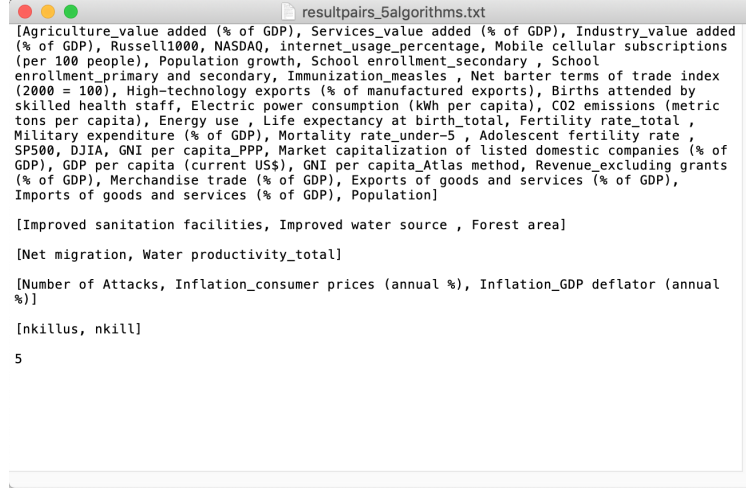
Taken as a whole, the bulk of existing research suggests that (1) countries with better functioning banks and markets grow faster; (2) simultaneity bias does not seem to drive these conclusions, and (3) better functioning financial systems ease the external financing constraints that impede firm and industrial expansion, suggesting that this is one mechanism through which financial development matters for growth.[9]

Based on the result of their survey, the new reference data we added following exploring how our majority vote works following three variables:personal spend; federal tax; and Stock Market Capitalization to GDP for World. In this case, our reference model of economic growth should include the list of variables in Table 5. In this table, the red color variables represent the stock market and stock market datasets. The green color represents the financial. The variable with the black text represent other influences for GDP growth, such as high technique or agriculture and industrial improvement. This table contains the known patterns from the economic field. We used it as reference model; if we added more variables, the cause and reasons would still exist as the data pool grows.

Table 5: Table for reference model's variables table

Net barter terms of trade index
Gross capital formation
SP500
High technology exports
Russell1000
DJIA
GDP per capita
Merchandise trade
Revenue
NASDAQ
Energy use
Inflation
GNI per capita
Imports of goods and services
Exports of goods and services
Market capitalization of listed domestic companies
Industry
Agriculture
GNI per capita
GDP growth

Figure 38 and Figure 39 list all the nodes inside the subgraphs. We can figure out that the second loop in Figure 39 perfectly matched our reference model for economy change.



```

resultpairs_5algorithms.txt
[Agriculture_value added (% of GDP), Services_value added (% of GDP), Industry_value added (% of GDP), Russell1000, NASDAQ, internet_usage_percentage, Mobile cellular subscriptions (per 100 people), Population growth, School enrollment_secondary , School enrollment_primary and secondary, Immunization_measles , Net barter terms of trade index (2000 = 100), High-technology exports (% of manufactured exports), Births attended by skilled health staff, Electric power consumption (kWh per capita), CO2 emissions (metric tons per capita), Energy use , Life expectancy at birth_total, Fertility rate_total , Military expenditure (% of GDP), Mortality rate_under-5 , Adolescent fertility rate , SP500, DJIA, GNI per capita_PPP, Market capitalization of listed domestic companies (% of GDP), GDP per capita (current US$), GNI per capita_Atlas method, Revenue_excluding grants (% of GDP), Merchandise trade (% of GDP), Exports of goods and services (% of GDP), Imports of goods and services (% of GDP), Population]

[Improved sanitation facilities, Improved water source , Forest area]

[Net migration, Water productivity_total]

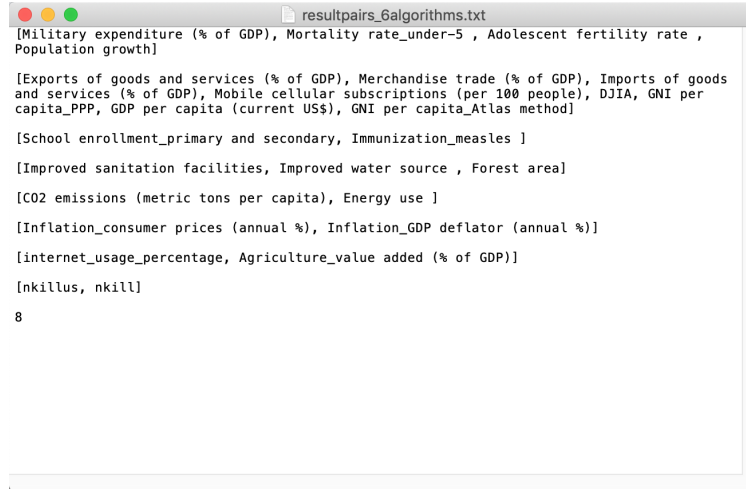
[Number of Attacks, Inflation_consumer prices (annual %), Inflation_GDP deflator (annual %)]

[nkillus, nkill]

5

```

Figure 38: Nodes for five algorithms support the result of Majority Vote



```

resultpairs_6algorithms.txt
[Military expenditure (% of GDP), Mortality rate_under-5 , Adolescent fertility rate , Population growth]

[Exports of goods and services (% of GDP), Merchandise trade (% of GDP), Imports of goods and services (% of GDP), Mobile cellular subscriptions (per 100 people), DJIA, GNI per capita_PPP, GDP per capita (current US$), GNI per capita_Atlas method]

[School enrollment_primary and secondary, Immunization_measles ]

[Improved sanitation facilities, Improved water source , Forest area]

[CO2 emissions (metric tons per capita), Energy use ]

[Inflation_consumer prices (annual %), Inflation_GDP deflator (annual %)]

[internet_usage_percentage, Agriculture_value added (% of GDP)]

[nkillus, nkill]

8

```

Figure 39: Nodes for six algorithms support the result of Majority Vote

### 6.1.2 Jaccard index

To evaluate the behavior of the support matrix, we used Jaccard index. The Jaccard index, also known as Intersection over Union and the Jaccard similarity coefficient, is a statis-

Table 6: Table for reference majority vote results.

algorithm support	pairs
0	1461
1	423
2	91
3	54
4	34
5	26
6	27

tic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.[18]The function of the Jaccard index is:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (6.1)$$

In function 6.1, the value is 0 when the two sets are disjoint, 1 when two sets are equal, and strictly between 0 and 1 when they are otherwise. Two sets are more similar when their Jaccard index is closer to 1. Here we defined A as our reference pairs and B as our majority vote value. Table 5 shows the result pairs for the reference majority vote table. After selecting the non-normal distributed variables, we chose 46 variables with 2116 pairs. 27 pairs were supported by all algorithms. Taking the 27 pairs for further analyses, we found them in 8 DAGs. We calculated the Jaccard index of the following list of variables together with table 4 as our reliability of our majority vote algorithms support.

- High technology exports, Forest area, Improved sanitation facilities, Improved water source, Mobile cellular subscriptions, GNI per capita
- Military expenditure, Mortality rate

Table 7: Table for reference majority vote results.

Numbers of supporting algorithms	4	5	6
Existence of the new variables	3	3	2
Strict direction of same pairs	20	15	7
Total pairs for reference model	87	53	27
Number of subgraph(new/original)	4/7	8/5	8/8
Jaccard indexs	22.94%	28.3%	25.92%
Total of same variable/reference model nodes	51.16%	74.19%	52.38%

- Inflation consumer prices, Inflation GDP deflator
- Services value added, Industry value added
- Merchandise trade, Imports of goods and services, Exports of goods and services
- Federal tax, Market capitalization of listed domestic companies
- Personal spend, GDP per capita
- DJIA, SP500

Table 7 shows the results for our support matrix and the reliability of possible causal discovery. Majority vote in discovering causal relationship is a good way to avoid the unreliability single algorithm may have. However, it does not mean the larger the majority vote is, the better the result is. Rather, it is based on the fact that some of the algorithms are more sensitive in dealing with small datasets(PC), while others like FCI will give the latent value. But majority vote behaves well in finding the potential related variables. Our results were supported by a reference model.

## 7.0 CONCLUSIONS

In this thesis, we explored open-source websites and found a way to store and merge data. We learned how to use Influx DB to realize building a database and how to visualize the data in Grafana. We built a prototype of a social weather system. Our website could be used as a visualization for time-series databases and exploration for general trends among the variables.

For merging the datasets together in time series data, setting same timestamps is important since no two entries will have the same timestamp and the same tag. The old data will be replaced by new one if we try to push new data with same timestamps and same tag in influx db. The only way is that to either change the tag value or change the timestamps.

Based on the fact we only have acknowledged causal relationships in a specific subject, as the datasets grow larger, the variables in one area may have other factors of influence in other subjects, and causal discovery is more useful in dealing with this situation.

We explored the causal discovery algorithms and applied them to discover relationships between social variables. In our experiment, our support matrix works well in finding related nodes. The reliability for five algorithms in finding nodes with real causality is 74.19%, but goes to 52.38% when it rises to six algorithms support the results. The results reflected the FCI and GFCI algorithm: hence variable A and variable B don't directly have causality  $\text{pair}[A,B] = [0,0]$ . The result of majority vote becomes smaller while there is six algorithm support.

Causal discovery can help reduce the range of similar trend variables for future use. We used the voting approach to discover strong relationships. Majority vote in discovery of causal relationship is a good way to avoid the unreliability that single algorithm may have. But it does not mean that the larger the majority vote is, the better the result is.

Rather based on the fact that some of the algorithms are more sensitive in dealing with small datasets(PC), while others like FCI will give the latent value. But majority vote behaves well in finding the potential related variables. Our results were supported by a reference model. By using the subgraph we found a new way to decrease the target dataset when doing the future work.

## BIBLIOGRAPHY

- [1] Cooley, R., Mobasher, B., & Srivastava, J. (1997, November). Web mining: Information and pattern discovery on the world wide web. *In Tools with Artificial Intelligence*, 1997. Proceedings., Ninth IEEE International Conference on (pp. 558-567). IEEE.
- [2] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *Acm Sigkdd Explorations Newsletter*, 12, 12-23.
- [3] Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, prediction, and search*. MIT press.
- [4] J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):18731896, 2008.
- [5] Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3), 199-236.
- [6] Colombo, D., Maathuis, M. H., Kalisch, M., & Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 294-321.
- [7] Coumans, Vincent, Tom Claassen, and Sebastiaan Terwijn. *Causal Discovery Algorithms and Real World Systems*. (2017).
- [8] Aghion, P., & Howitt, P. W. (2008). *The economics of growth*, MIT press.
- [9] Levine, R. (2005). *Finance and growth: theory and evidence. Handbook of economic growth*, 1, 865-934.
- [10] Boyer, R. S.; Moore, J S. (1991), "MJRTY - A Fast Majority Vote Algorithm", in Boyer, R. S., *Automated Reasoning: Essays in Honor of Woody Bledsoe*, Automated Reasoning Series, Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 105-117, Originally published as a technical report in 1981.
- [11] Chickering, D. M. (2002). Optimal structure identification with greedy search, *Journal of machine learning research*, 3(Nov), 507-554.



- [12] Pearl, J. (2003). Causality: models, reasoning, and inference. *Econometric Theory*, 19(675-685), 46.
- [13] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396), 945-960.
- [14] Andrews B, Ramsey J, Cooper, GF. Scoring Bayesian networks of mixed variables. *International Journal of Data Science and Analytics*, 9 December 2017
- [15] Raftery, A. E. (1999). Bayes factors and BIC: Comment on A critique of the Bayesian information criterion for model selection. *Sociological Methods & Research*, 27(3), 411-427.
- [16] Manning, P., Francois, P., Hoyer, D., & Zadorozhny, V. (2017). Collaborative Historical Information Analysis. *Comprehensive Geographic Information Systems*, Pages 119-144.
- [17] Wu, X., Guo, C. X., & Cao, Y. J. (2005). A new fault diagnosis approach of power system based on Bayesian network and temporal order information. *Proceedings-Chinese Society of Electrical Engineering*, 25(13), 14.
- [18] Yang, L., Zhi, Y., Wei, T., Yu, S., & Ma, J. (2019). Inference attack in Android Activity based on program fingerprint. *Journal of Network and Computer Applications*, 127, 92-106.